

Latest Developments of the Linked Thesaurus Framework for the Environment (LusTRE)

Andreas Abecker, Karsten Schnitter, Roman Wössner
Disy Informationssysteme GmbH

Karlsruhe, Germany
firstname.lastname@disy.net

Riccardo Albertoni, Monica de Martino, Paola Podestà
Istituto di Matematica Applicata e Tecnologie Informatiche,
Consiglio Nazionale delle Ricerche (CNR-IMATI)
Genova, Italy
firstname.lastname@ge.imati.cnr.it

Abstract— We present a solution for inter-linking domain thesauri and other controlled vocabularies represented as Linked Data content within the LusTRE framework. We discuss how such inter-linked content can help users to easier and better express and use thesauri and controlled vocabularies for metadata work within Spatial Data Infrastructures, at the concrete example of an INSPIRE-driven environmental use case. We present the Thesaurus Exploitation services designed and implemented for LusTRE and their integration with existing metadata editors and SDI geoportals.

Keywords—thesaurus management, thesaurus exploitation service, thesaurus inter-linking, linked open data, SKOS, metadata management, INSPIRE implementation

I. INTRODUCTION

The EU-FP7 CIP-PSP project eENVplus [3] aims at developing a complete ecosystem of software services for comprehensively supporting all aspects of interoperable environmental data in the Web. The software services are developed for and tested in ten INSPIRE and SEIS pilot implementation test-beds. INSPIRE¹ is the European Directive which pledges the EU member states to establish Web-based and service-oriented software infrastructure for spatial information to support Community environmental policies, and policies or activities which may have an impact on the environment. INSPIRE addresses 34 data themes (like cadastral parcels, hydrography, protected sites, agricultural+aquaculture facilities, sea regions, natural risk zones, etc.) needed for environmental applications, with key components specified through technical implementing rules. SEIS², the Shared Environmental Information System for Europe aims at better access to environment information for environment administration and politics, NGOs, academia and general public. The eENVplus pilot test-beds comprise, e.g., a cross-border pilot about everyday-life issues connected to environmental aspects in Czech Republic and Slovakia, a natural-areas INSPIRE-compliance toolbox in France, a cross-border pilot in Hungary and Slovakia about protected areas, a cross-border pilot in Italy and Slovenia about geological-map harmonization, etc.

For all the data published in such Spatial Data Infrastructures (SDIs) for the environment, *metadata* must be added

¹ <http://inspire.ec.europa.eu/>

² <http://ec.europa.eu/environment/archives/seis/>

which describe general aspects of the actual data such as the creator, their up-to-dateness, their geographical coverage or keywords about their subject-matter. In order to “standardize” such content descriptions, keywords for subjects are normally taken from pre-defined thesauri or vocabularies. However, such pre-defined “Knowledge Organization Systems (KOS)” are often made for domain experts and cover only very narrow knowledge areas (but very deeply), they are mostly not translated in all European languages, and they adopt different viewpoints for structuring their concepts. Hence, although the different KOSs logically overlap and are related, this is normally very difficult to oversee for an end user – which leads to limited interoperability of metadata written with respect to different KOSs. Hence, as part of the eENVplus project, a thesaurus framework (LusTRE - Linked Thesaurus Framework for the Environment) is being developed and exemplarily filled which allows inter-linking of different KOSs (domain thesauri) in order to achieve a broader domain coverage, more multilinguality, and novel cross-domain thesaurus services when using KOSs for creating or editing metadata or for searching. The LusTRE framework and its associated exploitation services (LusTRE-ES) are designed as a stand-alone software infrastructure which can be called in a service-oriented manner by other software tools such as metadata editors or catalogue services.

Intermediate versions of the LusTRE framework and its exploitation services have already been presented in prior papers and talks (see, e.g. [1] or [2]). Now, coming close to the end of the project, we sketch the almost final status of the software developments. The presentation in this paper is directed towards an audience with prior knowledge, in the LusTRE work, the topic of thesauri in INSPIRE implementation, or technical details of thesaurus and/or metadata management.

This paper is structured as follows: In the next Section II, an overview of LusTRE system components and their interplay is given, including some abstract use cases illustrating the potential benefits for end users working with the tool. In Section III, the current status of LusTRE content is sketched. Section IV lists the available LusTRE Exploitation Services, illustrates one concrete service in more detail and briefly reviews the LusTRE-ES implementation. Section V describes the LusTRE web interface and the explorative tool functionalities offered to users for manual inspection of inter-

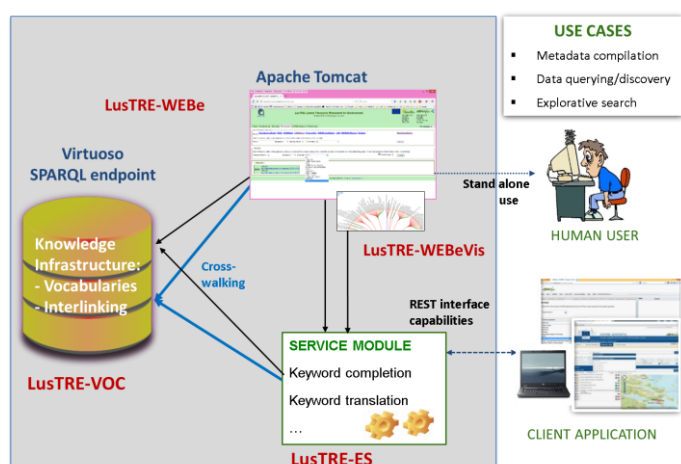


Figure 1: Overview of LusTRE Components

linked thesauri. Section VI mentions the tools which have already exemplarily been coupled to and enhanced by LusTRE-ES. Section VII concludes the paper with a short summary.

II. OVERVIEW OF LUSTRE COMPONENTS

Figure 1 illustrates the different elements of the LusTRE solution and their collaboration. The knowledge infrastructure of LusTRE is LusTRE-VOC. The LusTRE-VOC implementation is a Virtuoso instance. Virtuoso³ is a multi-purpose and multi-protocol data server which, among other functionalities, allows storing and querying RDF knowledge bases. RDF is the representation language of the Semantic Web for easily representing arbitrary knowledge structures. Virtuoso offers a SPARQL endpoint. SPARQL⁴ is the World Wide Web Consortium's (W3C) standard query language for RDF. The thesauri, ontologies, controlled vocabularies or code lists that we want to manage and link together in LusTRE are all different kinds of Knowledge Organization Systems. For such KOS, the "Simple Knowledge Organization System" (SKOS)⁵ has been proposed as the W3C standard way of representing in RDF. With SKOS, all thesaurus-relevant aspects can be expressed, such as concepts with different labels and translations of labels, broader and narrower relationships, more general relationships, etc. Further, the Semantic Web languages in general allow expressing that specific objects are identical ("matching statements"). Hence, our approach in LusTRE is to represent existing environmental KOS in the SKOS schema and inter-link them by matching statements. This general approach follows the Linked Open Data (LOD) paradigm pushed by the W3C since a couple of years.⁶ For doing so, all integrated thesauri must be expressed in SKOS format and all concepts represented by a dereferenceable URI. For some thesauri, such SKOS representations of existing KOSs have been created within the project and stored in LusTRE-VOC. But if a KOS is already published as SKOS following the LOD

³ <http://www.w3.org/wiki/VirtuosoUniversalServer>

⁴ <http://www.w3.org/TR/rdf-sparql-query/>

⁵ <http://www.w3.org/2004/02/skos/>

⁶ <http://www.w3.org/wiki/LinkedData>

principles, it can remain with the created and only the inter-linkings to such external sources must be represented in LusTRE-VOC. After the inter-linking of different environmental thematic vocabularies they can be accessed through the SPARQL endpoint as one virtual integrated linked data source.

In order to ease the use of the content contained in LusTRE, without knowing the technicalities of SPARQL and the underlying knowledge infrastructure, the LusTRE exploitation services (LusTRE-ES) have been implemented. LusTRE-ES, through a REST interface, offers a number of end-user oriented web services that can help with keyword management when working on one of the typical use cases: creating metadata for some web data resource in an SDI, querying or discovering web resources through metadata, or exploring the conceptual space spanned by resource metadata in order to browse the underlying knowledge, better understanding the application domain, etc. These LusTRE-ES services are implemented by a number of service modules that are coordinated by a central service manager. Concretely, such a service module takes a specific user request with the associated input parameters, formulated in terms appropriate for the end user functionality, and translates it into the appropriate SPARQL query to be evaluated against the LusTRE-VOC SPARQL endpoint. The service manager takes these SPARQL query/ies (sometimes, several service modules may be suitable for a given user request), hands them over to the Virtuoso server, collects the results, if necessary merges the results, and delivers the final result to the client. During the execution of the SPARQL query by the Virtuoso server, the cross-walking between different thesauri may happen – thus connecting different knowledge domains or bridging between different languages, etc.

Such clients which use the LusTRE-ES service interface to "pimp" their own functionalities with (linked) thesaurus knowledge, may, for instance, be metadata editors, catalogue services, or geoportals.

Besides this kind of "embedded" use of LusTRE, hidden behind the functionalities of another client software, a human user may also be interested in directly, manually exploring the inter-linked thesauri. To this end, the LusTRE-WEBE web interface has been built that allows, on one hand, manual browsing through the LusTRE vocabularies, and, on the other hand, calling the LusTRE-ES services directly. This web interface will also contain an enhanced visualization tool (LusTRE-WEBEVis) that facilitates graphical interaction with the thesaurus knowledge and illustrates the knowledge graphs and their interdependencies.

Abstract Use Cases.

In general, use cases for the provided functionalities can be sketched as follows:

- The LusTRE exploitation services can deliver to a search engine or to a metadata editor, the synonymous, broader, narrower or related terms of a keyword, taken from several inter-linked controlled vocabularies (KOSs) – which facilitates cross-domain, cross-corpus, cross-boundary and multilingual search as well as consistent and complete metadata annotation.

- For a user who is less experienced in some domain and who has to create or maintain metadata, the visualization service as well as visual browsing through conceptual spaces of several inter-linked thesauri, together with term explanations and term translations, can help to better understand and use the provided controlled vocabularies.
- The automated cross-walking between terms from different inter-linked KOSs supports easier working beyond the scope and limitations of a single vocabulary alone.

III. CONSOLIDATION OF THE LUSTRE CONTENT

During the last year, the set of included KOSs (ontologies, thesauri, code lists) has been extended (see Figure 2), as well as their inter-linking. In the meanwhile, the content in the LusTRE knowledge infrastructure covers all INSPIRE data themes. In particular, extensive and inter-linked vocabularies are available for the following data themes:

- PS (Protected Sites).
- BR (Biogeographical regions).
- HB (Habitat & Biotope).
- SD (Species Distribution).
- EF (Environmental Monitoring Facilities).
- SO (Soil).
- AQ (Air Quality).

Besides that, good language coverage and a good level of multilingualism have been reached.

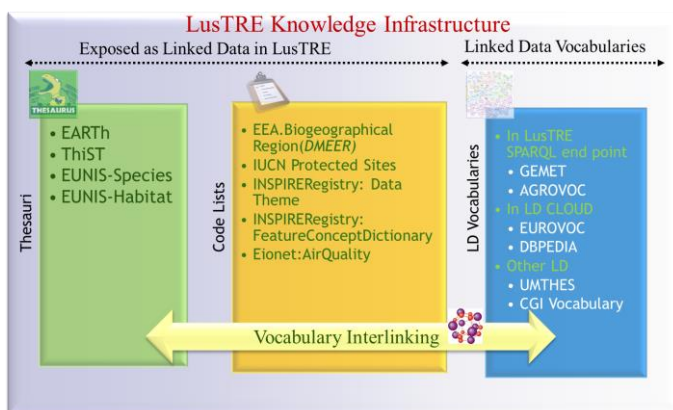


Figure 2: Overview of Controlled Vocabularies Currently Included in LusTRE

IV. AVAILABLE LUSTRE EXPLOITATION SERVICES

In the last year, the software services for using the thesaurus knowledge in metadata creation and information search (specified in [1]), have been fully implemented, extensively tested, and – based on user feedback – slightly extended and refined. Two additional parameters have been included which allow to better control the extent to which the conceptual space spanned by the inter-linked concepts in different controlled vocabularies, is really explored in concrete

queries to the LusTRE (for instance, when asking for synonymous, broader, narrower, or related terms). One new parameter allows focusing only on a specific subset of available thesauri when evaluating a request. The other new parameter allows switching on or off whether inter-thesaurus should really be used when evaluating a request (instead of restricting the search to concepts all contained in one specific thesaurus). The following services are offered to the end user or to any client application through the REST interface:

- **GetCapabilities:** Describes the services offered by this concrete LusTRE-ES instance. The response contains (1) service metadata; (2) operation descriptions for the implemented service modules, their operations and the parameters of these operations; and (3) descriptions of the thesauri exploited by this LusTRE-ES instance.
- **DescribeConcept:** Returns the SKOS descriptions from the associated thesauri by directly retrieving and handing over a SPARQL request with the given input keyword sent to the Virtuoso instance that stores the associated thesaurus knowledge.
- **SPARQL:** Similarly to **DescribeConcept**, this operation allows the user to run own SPARQL queries against the Virtuoso server's SPARQL endpoint.
- **GetSuggestions:** If the user starts to type some keyword and the first few keystrokes are sent to the LusTRE-ES, this operation will offer potential keyword completions.
- **GetRelatives:** Retrieves all concepts related to the input concept through a broader, related or narrower relation, if specified by the cross-walking parameter, also taking into account inter-thesaurus links.
- **GetSynonyms:** Finds all synonymous keywords for a given concept. If specified by the cross-walking parameter, also taking into account exactMatch inter-thesaurus links between different KOSs.
- **GetTopmostConcepts:** Lists the topmost concepts for a given thesaurus.
- **ResolveThesaurus:** For a given concept, finds the thesaurus it is contained in.

Service Example in Detail.

As an example for the LusTRE-ES functionalities and its usage, consider the **GetSuggestions** operation. The **GetSuggestions** operation is used to find concepts which have a label starting with the provided keyword (i.e. a sequence of characters typed in by the user). Case is ignored. As labels currently `skos:prefLabel`⁷ and `skos:altLabel`⁸ are considered. When specific languages are given as input parameter, LusTRE-ES considers only labels of the given languages. Similarly, only requested thesauri or service modules are used. By default, *all* available languages, thesauri and service modules are used.

⁷ <http://www.w3.org/2004/02/skos/core#prefLabel>

⁸ <http://www.w3.org/2004/02/skos/core#altLabel>

The `GetSuggestions` operation supports the following parameters:

Parameter	Type	Mandatory	Description
<code>keyword</code>	String	required	Keyword as search criteria, start of label
<code>maxCount</code>	Integer	optional	Maximum number of results
<code>thesauri</code>	String	optional	Comma-separated list of thesaurus-URIs
<code>service</code>	String	optional	Comma-separated list of <code>serviceIds</code> as given by <code>GetCapabilities</code>
<code>languages</code>	String	optional	Comma-separated list of languages based on ISO-639-1
<code>cross-walking</code>	Boolean	optional	Default is false, see below
<code>source</code>	Boolean	optional	Default is false, when true result contains thesaurus source

The `GetSuggestions` operation can be enhanced by specifying the `cross-walking` parameter. With `cross-walking=false`, this generally leads to a smaller but more focused result set. With `cross-walking=true`, LusTRE-ES will take into account the concept URI and consider all concepts linked by `skos:exactMatch`⁹ as the same concept.

Service Implementation.

The LusTRE-ES implementation code is organized in three separate components, the REST endpoints, the ServiceManager and the ServiceModules. Entry point for any call is the LusTRE-ES REST endpoint. It decodes the HTTP request and creates a suitable query for the ServiceManager. The ServiceManager delegates the query to all suitable ServiceModules. The modules return executable units of work that will retrieve the query responses from the SPARQL endpoint of LusTRE. These units of work are executed by the ServiceManager. After merging the results of the different executions, the ServiceManager returns the results to the REST endpoint. After creating the proper JSON response it is sent with the HTTP response. This architecture separates the REST endpoints from the Service Modules which contain the actual logic on how to answer the incoming requests. Furthermore it enables the possibility to retrieve results from different modules.

V. LUSTRE WEB INTERFACE AND EXPLORATION TOOL

In the last year, the LusTRE explorative Web GUI (LusTRE-WEBe) has been completely re-implemented in order to be more efficient and more user-friendly (see

⁹ <http://www.w3.org/2004/02/skos/core#exactMatch>

<http://linkeddata.ge.imati.cnr.it/>). The user interface is divided in different sections, including:

- The *home tab* in which you find a short description of LusTRE.
- The *vocabularies tab* in which you can find all the information about which thesauri are included in LusTRE, the vocabulary versions, licenses and provenance.
- The *service tab*, providing the concise technical documentation required by third parties that want to integrate and exploit the LusTRE-ES services in their clients.
- The *indicators tab* which provides useful information, such as access statistics, linguistic coverage, etc.
- The *reference tab* in which you find information about publications pertaining to LusTRE.
- The *exploration tool* to search and browse the knowledge infrastructure of LusTRE.

The *exploration tool* has been implemented to make accessible to the user the knowledge infrastructure in LusTRE for browsing and searching for concepts in the available vocabularies. Its functionalities are:

- (i) To perform content negotiation compliant to Linked Open Data requirements and W3C best practices.
- (ii) To perform human readable browsing of the LusTRE content.
- (iii) To offer smart services to aid the users to better exploit the LusTRE content – such as “translation” by cross-walking between vocabularies.

A user can access to the LusTRE content via the exploration Tab which provides different entry points:

1. Browsing the Knowledge Infrastructure by clicking on the list of vocabularies included in LusTRE (Search&Browse concept window).
2. Search for a concept by starting typing a keyword (Search&Browse concept window).
3. Translation of a concept available in LusTRE (Translation window).

The Search&Browse activity can be useful when searching the keywords for annotating data. The user can search for a keyword, indicating a language and indicating whether she is interested in searching into the whole inter-linked framework or only in a specific vocabulary. The tool returns all the concepts matching the keyword and the respective vocabulary source. Then a search-refinement activity may be further performed by selecting one of these concepts. By clicking on a concept, the user gets all the information related to it:

- The concept’s description page. The concept page includes useful and detailed information about the concepts, e.g., the concept URI which can be used by linked-data applications to get the machine-interpretable serialization of the concept.
- The lexical information of the concept such as the definition, translations and synonyms.
- The semantic relation with other concepts such as: Broader, narrower and related concepts. They can be exploited to make coarser or finer the keywords to be

exploited (i) in search and annotation tasks carried out in a client application which might not be yet integrated with the LusTRE-ES, or (ii) to discover related concepts which might be useful to find out, or (iii) by a user who wants to discover more about concepts and terminologies pertaining to specific domains.

For example, a user could decide to expand his conceptual search by exploring the broader term learning what it means and discovering that it is somehow related to other concepts which might be of interest, too.

LusTRE translation

LusTRE as a framework of inter-linked vocabularies, can not only be exploited stand-alone, but also jointly with others. We provide two practices for the vocabularies:

- The stand-alone use of a vocabulary which is basically the usual practice, similar to what is provided by the web sites of vocabularies such as GEMET and Agrovoc. The LusTRE contribution when considering the stand-alone use of a vocabulary is basically the publication as linked data of vocabularies which were not yet published in the LOD, such as THIS_t and EARTH.
- The joint exploitation of vocabularies within LusTRE which relies on the notion of cross-walking. The cross-walking is when LusTRE is taking advantage of inter-linking among vocabularies in order to enrich the set of information associated to concepts in one vocabulary.

Cross-walking brings an advantage in the translation as concepts which are not available in one language could be imported from other inter-linked vocabularies. For example: Let us suppose the user is interested in translating the word “soil” by the specific vocabulary EARTH which provides only Italian and English translations. First, the tool provides the set of concepts matching with “soil”, and the user selects the concept to be translated. Without cross-walking the tool provides synonyms available in EARTH which are only in English and Italian. Let’s repeat the search exploiting LusTRE as framework by checking the cross-walking functionality which returns not only translations and synonyms in EARTH, but also all those in EARTH’s inter-linked vocabularies. Then, the tool returns much more translations than before.

VI. INTEGRATION

In order to demonstrate and test the LusTRE usefulness and usability from the end-user perspective, the LusTRE exploitation services have been exemplarily integrated into two end user tools:

- The EUOSME metadata editor has been extended by LusTRE exploitation services to include novel features such as automated keyword completion, semantic search on keyword concepts and synonyms, or multi-language support (<http://showcase.eenvplus.eu/client/editor.htm>).
- The developer version of the JRC INSPIRE Geoportal has been prototypically extended by LusTRE exploitation services, e.g., to enable semantic search.

The work presented in this paper has been developed within the FP7 CIP-PSP pilot project eENVplus (eEnvironmental services for advanced applications within INSPIRE) partially funded by the European Commission under contract no. 325232.

VII. SUMMARY

In the last year, the work on LusTRE and the LusTRE exploitation services (LusTRE-ES) have made considerable progress. The content has been extended and was deeper interwoven with the Linked Open Data cloud. The LusTRE-ES implementation has been consolidated and tested and was extended by two parameters, the most important of which is the cross-walking parameter which allows inclusion of cross-vocabulary links in all operations. LusTRE has been given a more user-friendly and more functional GUI (LusTRE-WEBe) and is being extended by a visualization component (LusTRE-WEBeVis; not discussed in this paper; not yet fully implemented). The integration of LusTRE-ES with the EUOSME metadata editor and with the experimental INSPIRE Geoportal (developer version) allows for deep and broad experimentation and validation of LusTRE and TFES in the last project year by the pilot users. The technical interoperability has already been proven, but real end-user feedback about usability and perceived usefulness is still missing.

In general, we consider LOD principles and technologies a very promising approach to build federated thesaurus infrastructures, leaving the development and maintenance of thesauri in a decentral manner with their originators, but leverage the individual thesauri’s value by their connection and inter-linkage. In total, such an approach may globally save resources and locally increase value. However, how important the problems addressed really are for practitioners, must be shown by practice. Regarding the middleware design and base technologies used in and developed for the LusTRE implementation, we see easy ways for future extensions and take-up. In particular, we are considering the use of such an approach for flexibly adding powerful thesaurus functionalities to our commercial metadata framework Disy Preludio [4].¹⁰

REFERENCES

- [1] A. Abecker, R. Albertoni, C. Heidmann, M. de Martino and R. Wössner, “Using interlinked thesauri for INSPIRE-compliant metadata management,” in 28th Int. Conf. on Informatics for Environmental Protection: ICT for Energy Efficiency - EnviroInfo 2014, J. Marx Gómez et al, eds., pp. 509-516, 2014.
- [2] A. Abecker, R. Albertoni, C. Cipolloni, M. de Martino, Y. Moradiafkan and R. Wössner, “Software services exploiting the eENVplus framework of interlinked thesauri for metadata management,” in INSPIRE conference 2014: INSPIRE for good governance, Aalborg, Denmark, June 2014.
- [3] C. Attardo and G. Saio, “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, 2013.
- [4] A. Abecker, C. Heidmann, W. Kazakos, R. Nedkov, R. Wössner, “The Preludio metadata management system for comprehensive support of INSPIRE compliant metadata,” in INSPIRE conference 2014: INSPIRE for good governance, Aalborg, Denmark, June 2014.

¹⁰ <http://www.disy.net/produkte/preludio.html>