Using Interlinked Thesauri for INSPIRE-Compliant Metadata Management

Andreas Abecker[1], Riccardo Albertoni[2], Carsten Heidmann[1],
Monica De Martino[2] and Roman Wössner[1]

Abstract

As part of the eENVplus project about infrastructures for the INSPIRE implementation, a Thesaurus Framework (LusTRE) is being developed which allows to interlink different environmental domain thesauri and offers access to them as one virtual integrated linked data source – which shall support better metadata compilation and metadata discovery for describing and finding INSPIRE data and services. Further, a Web Service infrastructure is being provided (the LusTRE Exploitation Services, LusTRE-ES) that allows to make optimum use of the knowledge contained in LusTRE for improving existing metadata tools. In this paper, we give an overview of LusTRE and present important aspects of the LusTRE-ES: its architecture and design principles, the REST interface, the list of services / service modules under development, and the way how these services shall improve existing metadata tools.

1. Overview

The EU-FP7 CIP-PSP pilot project eENVplus ("eEnvironmental services for advanced applications within INSPIRE") develops – within ten INSPIRE/SEIS pilot-implementation testbeds – an ecosystem of software services for comprehensively supporting all aspects of interoperable environmental information services. As part of eENVplus, a thesaurus framework (LusTRE) is being developed and exemplarily filled which allows interlinking of different domain thesauri. In this paper, we present the eENVplus thesaurus exploitation services: Web services exploiting the domain knowledge contained in the interlinked thesauri for the purposes of metadata creation and maintenance and for resource discovery. The services can be used by different tools such as metadata editors, metadata validators, CSWs, information portals or specific application solutions. Some exemplary use cases:

- The eENVplus thesaurus exploitation services can deliver to a search engine synonymous, broader, narrower or related terms from several interlinked thesauri or controlled vocabularies which facilitates cross-domain, cross-corpus, cross-boundary and multilingual search.

- 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 interlinked thesauri together with term explanations and term translations can help to better understand and use the provided thesauri, controlled vocabularies and code lists.

- The automated cross-walking service between terms from different interlinked controlled vocabularies supports easier working beyond the scope and limitations of a single thesaurus or controlled vocabulary alone.

The paper is structured as follows: In Section 2, we sketch the eENVplus Linked Thesaurus Framework for the Environment (LusTRE). Section 3 is about the LusTRE Exploitation Services, showing the software architecture, the REST interface and the list of exploitation service modules under development. Section 4 is about how to use LusTRE-ES and explains use cases and application

[1] Disy Informationssysteme GmbH, 76131 Karlsruhe, Germany, firstname.lastname@disy.net

[2] Consiglio Nazionale delle Ricerche, Istituto di Matematica Applicata e Tecnologie Informatiche (CNR-IMATI), 16149 Genova, Italy, lastname@ge.imati.cnr.it

scenarios with the help of some application mock-ups. Section 5 concludes the paper with some information about commercial use and about the current implementation status.

2. The Linked Thesaurus Framework for the Environment (LusTRE)

Although different directives (e.g., INSPIRE - Infrastructure for Spatial Information in the European Community) and policy communications (e.g., SEIS - Shared Environmental Information System) have been launched at European-scale with the objective of improving the management of heterogeneous environmental data sources, an effective sharing of these resources is still part of the desiderata, also due to the intrinsic multicultural, multilingual and multidisciplinary nature of the environmental domain. Terminological resources such as thesauri are widely employed as common ground enabling communication among the different communities working in environment-related domains: they allow users to share and agree upon scientific/technical terms in the target domain and to express them in multiple languages. In the recent years several thesauri have been deployed by different communities having a large spectrum of competencies. They have been created embodying different points of view based on different ways of conceptualization. Their development reflects different scopes and implies quite a range of levels of abstraction and detail. All these thesauri are precious and their reusability is pivotal within a Spatial Data Infrastructure (SDI) for providing more homogeneity in data description (metadata compilation) and data discovery.

In eENVplus, we are developing a multilingual, pan-European infrastructure of Knowledge-Organization Systems (KOS: thesauri, taxonomies, ontologies, ...) for the environment considered as a set of interlinked/interoperable terminologies distributed on the Web together with the the services needed for their management and usage. The LusTRE infrastructure addresses the needs of different user communities in sharing digital information at cross-border level, i.e. concept interoperability and concept-availability in multiple languages as it is needed in metadata compilation and information discovery.

In particular, we are building a software framework that allows combining existing thesauri to support the management of environmental resources. It considers the heterogeneity in scope and levels of abstraction of environmental thesauri as an asset when managing environmental data, so it exploits linked data best practices [3, 4] to provide a multi-thesaurus solution for INSPIRE data themes related to the environment. The LusTRE framework extends the common thesaurus framework for Nature Conservation [5] resulting from the NatureSDIplus project (http://www.nature-sdi.eu/). It is based on existing KOSs encoded in the SKOS ontology (Simple Knowledge Organization System) and represented in the RDF data model (Resource Description Framework). LusTRE provides an integrated view on different available terminologies. It is an open environment where it is possible to add, assemble and share in a frame general-purpose KOS as well as domain-specific terminologies. The framework has to be a dynamic environment in the sense that a new KOS may be added or a thesaurus within the framework can be extended easily. For this purpose, the following requirements were considered when building LusTRE:

- **Modularity**. Each KOS should be intended as a module plugged in the set of thesauri included in the framework. In particular, modularity should be preserved in order to eventually include future updates for existing terminologies.
- **Openness.** Each KOS should be easily extendable, in order to possibly add (as separated modules) new concepts and terms keeping separated the original terminology/thesaurus.
- **Exploitability.** Each KOS should be encoded in a standard and flexible format, in order to encourage the adoption and the eventual enrichment from third party system.
- **Interlinking.** Terms and concepts in existing KOS should be interlinked in order to harmonize the term usage from a multicultural point of view.

Altogether, the technology in the Semantic Web field aids to meet these requirements: Hence, the standard model SKOS (Simple Knowledge Organization System) is employed to encode each KOS and Linked Data principles are employed to expose, share, and connect thesauri via Uniform Resource Identifiers (URI) that can be looked up on the Web. In particular, the resource translation to SKOS enables the modularity property, whereas the resource accessibility according to linked data is a key aspect to meet the requirements of openness and exploitability.

At this stage of development, LusTRE is already available as linked data[3], it already includes differrent SKOS/RDF resources such as the **Thist** thesaurus for geology, the **EUNIS** Species and Habitat types, the Digital map on EU Ecological Regions (**DMEER**) and the Environmental Application Reference Thesaurus (**EARTh**) which is adopted as backbone thesaurus in the framework. Besides, it includes *intra-thesaurus interlinks* between Habitats and Species, DMEER and EARTh, as well as an *inter-thesaurus linkset* from EUNIS Species to official EUNIS species provided by the European Environmental Agency, and from EARTh to **GEMET**. Further inter-thesaurus equivalences have been added recently [1]. GEMET's outgoing links to **AGROVOC**, **EUROVOC**, **DBpedia** and **UMTHES** have been imported into EARTh by working out GEMET's *skos:exactMatch* relationships. Unfortunately, links obtained by this procedure only pertain to the subset of concepts that EARTh shares with GEMET. In order to complement that set and find out a more complete connection among EARTh and GEMET's linked datasets, a two-step process has been put in place: First, SILK[4] has been applied to discover new links, then the SILK results have been validated by experts in order to verify the accuracy of the links and to identify the most suitable types of interlinking property (i.e., *skos:exactMatch* or *skos:closeMatch*). The joint exploitation of *skos:exactMatch* transitive closure and the manually validated SILK link discovery have almost triplicated the number of outgoing links available with respect to the previous EARTh releases. In particular, about 7171 links have been discovered relying on the transitive closure, 465 have been generated deploying SILK. This new release paves the way for a combined exploitation of LusTRE with GEMET, AGROVOC, EUROVOC, DBpedia and UMTHES (about 33% of the EARTh concepts have a link to other thesauri) enabling LusTRE adopters in taking advantage of their respecttive strengths and complementarities.

3. An Architecture for Exploiting LusTRE Knowledge

3.1. Architecture of LusTRE Exploitation Server

Figure 1 illustrates the overall idea of how to make use of the thesaurus knowledge provided by LusTRE in order to improve metadata compilation, harmonization and data discovery, e.g. in geo portals: One one hand, we have end users, who typically create, edit or search metadata which describe spatial data sources or spatial data services. We cannot expect that all users should switch to a metadata management tool (geodata portal, CSW tool, etc.) provided by eENVplus; instead, we offer highly configurable services with standards-compliant interfaces and, in that way, allow all users to extend their existing client software with functionalities exploiting the LusTRE knowledge. This means, "client" in the figure below could be any existing (or new) tool for creating, processing or searching geo metadata. On the other hand, we have the different environmental thesauri which are – through the functionalities of the LusTRE Thesaurus Framework – bound together to a large, multilingual and multidisciplinary conceptual space. The knowledge contained in this conceptual space can, by technicians, be accessed through LusTRE's SPARQL endpoint. However, instead of directly addressing this SPARQL endpoint, the LusTRE

[3] http://linkeddata.ge.imati.cnr.it:2020/

[4] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

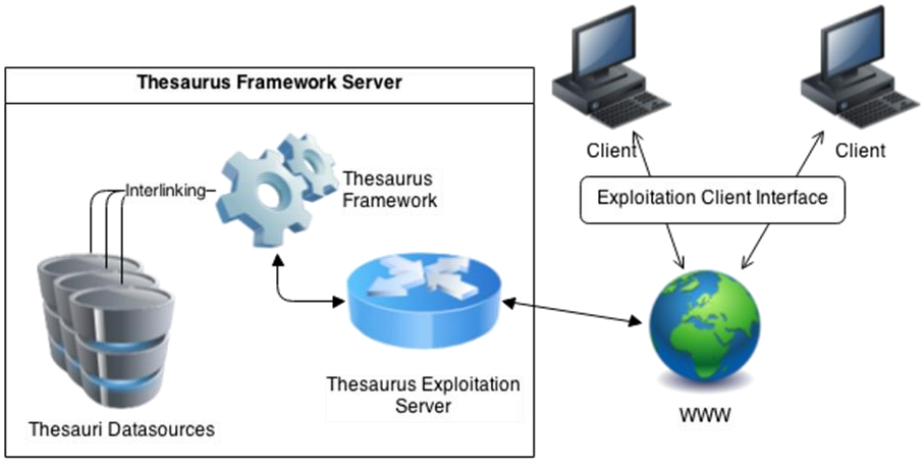

Figure 1: Overall Usage Scenario for Interlinked Thesauri

Exploitation Server (LusTRE-ES, Figure 2) manages all interactions with clients.

For better expandability and flexibility, the LusTRE-ES is designed as shown in Figure 2: It consists of an (i) HTTP Server that offers a REST interface; (ii) a Service Manager; and (iii) a non-empty set of Service Modules. The main task of the **Service Manager** is to delegate requests from the HTTP server to suitable Service Modules based on their exposed capabilities, to collect returned *Callables* from Service Modules, to orchestrate the execution of the *Callables* and route back the query results to the HTTP server.

The **Service Modules** provide the core functionalities of the system. They implement the various exploitation-functionality logics, practically by translating an end-user request in a (set of) SPARQL query/ies. The Service Modules are the only points of contact between the thesaurus framework back-end and the rest of the exploitation layer. The modules are like pluggable cartridges for the system engine, each of which enhances the behaviour of the system in certain domains or criteria. The architecture of the system allows these modules to be totally independent from each other, to focus on the one and only task they have: "perform a specialized query from the TF". For each request type supported by a Service Module, it must offer a separate interface that has a public method which returns a *Callable<T>* as 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

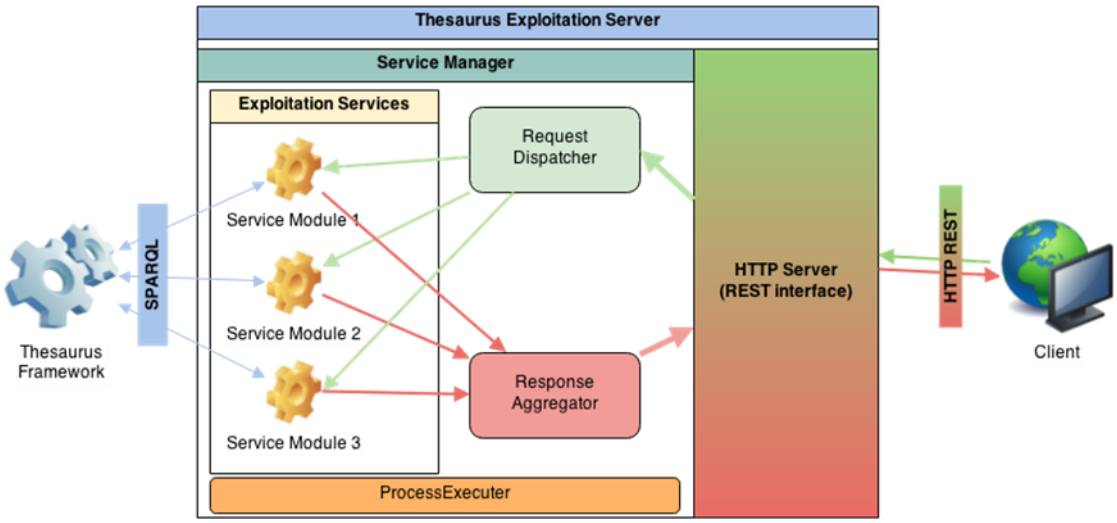

Figure 2: Components of LusTRE Thesaurus Framework Exploitation Server

rest of the logic, as well as any number of private methods and fields, can be freely defined and developed by the developer to address the task the service module have to carry out. Following our Service-Module design guideline, a uniform interface for all the Service Modules can be maintained. The advantage of such a design is that every Service Module includes *only* the solution logic for the task at hand, and the Service Manager can orchestrate the execution of the processes.

3.2. REST Interface Operations

The REST interface foresees one operation to be offered by all Service Modules and by the Service Manager, namely the *GetCapabilities* operation as the starting point for every client to gather the basic service information as well as a capabilities catalog as a guide to the available operations and resources. Further, there are two operations (*DescribeConcept*, *ResolveThesaurus*) that shall be answered by exactly one Service Module (see below). But the most often used operations for accessing thesaurus knowledge may be interpreted differently by different Service Modules.

- The purpose of the *GetSuggestions* operation is to suggest matching thesauri concepts upon receipt of a lexical keyword, or a part a lexical keyword. The matching logic and the actual criteria for matching the thesauri concepts depend on the active Service Modules in the system.
- The purpose of the *GetSynonyms* operation is to query for all concepts which are considered to be synonyms to the input concept. The equality logic and the actual criteria for finding the synonymous concepts depend on the active Service Modules in the system. For example, some module might deploy cross-walking through different thesauri in order to include inter-thesauri synonyms by considering also synonyms *(skos:altLabel)* and preferred labels *(skos:prefLabel)* coming from the interlinked *skos:conceptScheme*.
- The purpose of the *GetRelatives* operation is to query for all concepts related to the input thesaurus concept. The relationship logic and the actual criteria for finding the relative thesauri concepts to the given concept depend on the active Service Modules in the system.

3.3. List of LusTRE Exploitation Services Under Development

The following Table 1 includes the list of service modules and their respective capabilities currently being developed for the eENVplus project.

Service Module Name **Capabilities**	**Description of operation execution**
KeywordCompletion *GetSuggestions*	Given a series of characters as part of a keyword, this module will query the thesaurus framework and return a list of thesauri concepts. Selection logic: based on *prefLabel* and *altLabel*.
KeywordExplanation *DescribeConcept*	Given a concept URI, this module will return the complete SKOS concept record.
KeywordTranslation *GetSynonyms*	Given a concept URI, this module will return all the thesauri concepts that are considered synonyms and/or *exactMatch* to the selected concept in any given natural language available for the framework. The functionality includes also cross-walking through the interlinked thesauri to consider the mapped concepts from other thesauri.
ThesaurusResolution *ResolveThesaurus*	Given a concept URI, this module will return the originating thesaurus to meet TG Requirement 16 of the INSPIRE Metadata Implementing Rules - Technical Guidelines.
QueryReformulation	Given a concept URI, this module will return all the thesauri

GetSynonyms	concepts that are considered synonyms to the selected concept in any given natural language available for the framework. (also through cross-walking, respectively through thesauri interlinking)
AnnotationNormalization *GetSuggestions*	Given a concept URI, this module will return the best fitting thesauri concept, respect to the agreement of a literal text input and the preferred label of a concept.
SemanticExplorative Search *GetRelatives* *GetSynonyms* *DescribeConcept*	Given a concept URI, this module will return all related concepts and needed information, for example, to populate a semantic visualization.

Table 1: eENVplus Service Modules and Capabilities

Please note: (1) Some of the Service Modules offer a number of parameters to further describe the requests (for instance, a maximum number of suggested keywords, a limitation to specific natural languages or thesauri, etc.). (2) We usually allow to exchange so-called *KeywordObjects* as operation input and output that contain no only lexical terms, but also the concept URI, if available.

4. Use Cases for LusTRE Exploitation Services

Practically, we consider three basic use cases further explained in the subsections below.

4.1. Metadata Compilation

When an end user creates, harmonizes, or actualizes metadata, several LusTRE-ES services may be useful: Of course, **KeyWordCompletion** may always save time and reduce error possibilities, it may also increase the level of cross-indexer homogeneity. Using **KeywordExplanation** and **KeywordTranslation** may also help to better understand the offered indexing terms and their correct use – in particular, if one is not working in his mother language or preferred domain topic. Seeing broader, narrower and related terms, also from different thesauri (offered by **KeywordExplanation** or by **QueryReformulation**) may provide further help. A **KeywordValidation** module could even (semi-)automatically provide a „normal form“ for a given (set of) metadata record(s) that contains only preferred labels, etc. Such a validation procedure could also be run in batch-mode as a metadata transformation services applied to legacy metadata.

4.2. Metadata Discovery

Figure 3 illustrates a possible intended way of using the LusTRE-ES to improve resource discovery based on metadata. Assume a user typing some characters in the search field of his CSW, the LusTRE **KeywordCompletion** could offer a number of potential word extensions (1), maybe even

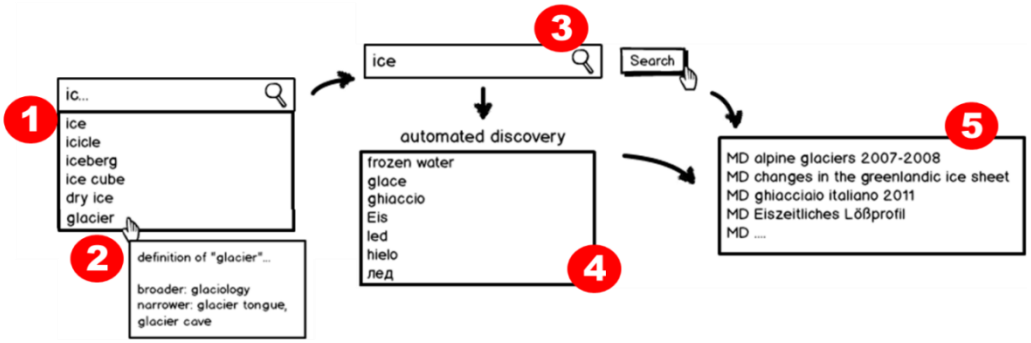

Figure 3: Mock-Up of Metadata Discovery Use Case

ones that only match at the conceptual, not at the lexical level (like glacier). Using **KeywordExplanation**, background knowledge about terms could be offered for tooltips (2). If the user decides for some search term (3) and starts the query, the **QueryReformulation** could provide an extended list of search terms such that also synonymous, translated or closely related terms could be included in the search (4), finally leading to a more complete result set (5).

4.3. Semantic Explorative Search

The third use case can be combined both with metadata creation and with resource discovery. Our so-called **SemanticExplorativeSearch** module provides the "conceptual neighbourhood" of a given keyword which allows building up an interactive, browsable concept map which helps an unexperienced user to better understand a concept in all its facets and relational contexts. Figure 4 illustrates the idea in a mock-up. Here, the term "Soil" is shown with its GEMET broader term and narrower terms as well as access to its definition and its translation. Further, there are navigations to related GEMET terms and one inter-thesauri link to the *skos:exactMatch*, namely "Soil" in EarTH – that has, e.g., completely different broader and narrower terms since it represents a different perspective on the topic.

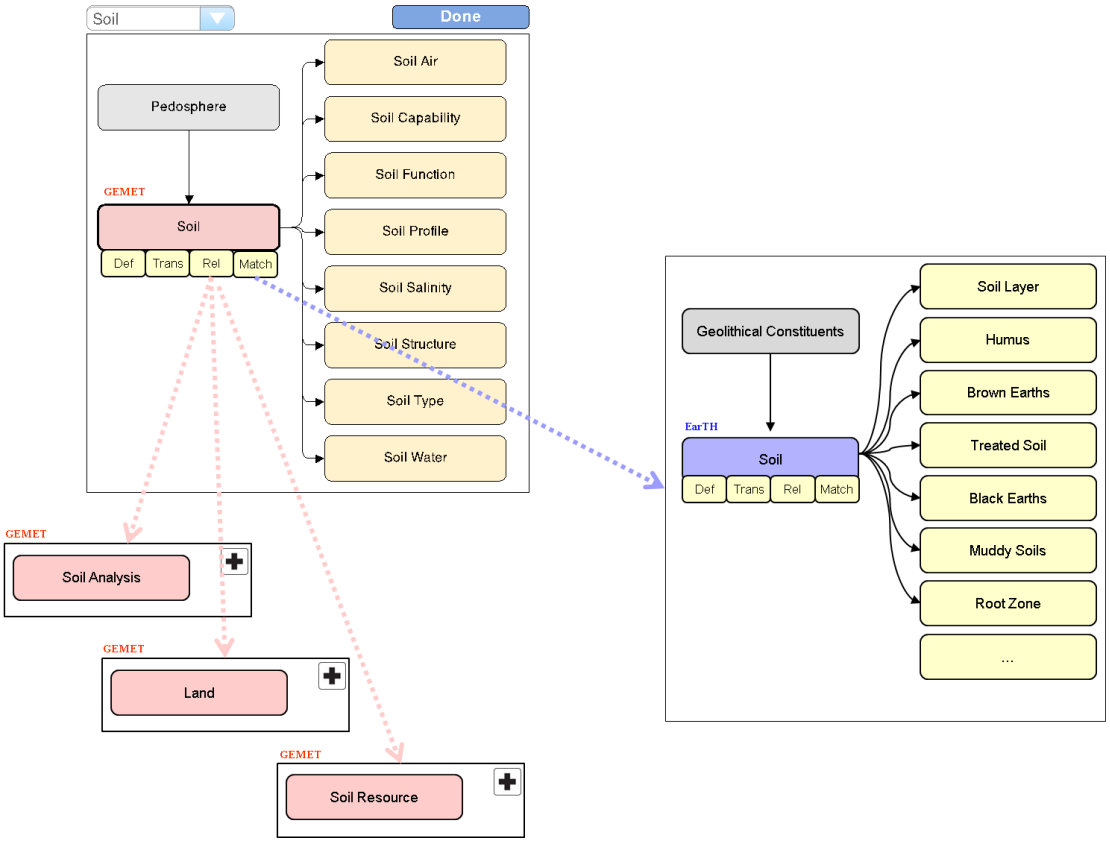

Figure 4: Mock-Up of Semantic Explorative Search Use Case

5. Conclusions

When writing this paper, the LusTRE is already publicly accessible. The LusTRE-ES architecture is running and will stepwisely be populated by Service Modules during summer and autumn 2014. The integration with different clients is being tested. The service usage will be prototypically implemented by integration with (1) the eENVplus metadata tool under development, (2) the JRC geoportal developer version and (3) with the commercial metadata management tool Disy Preludio.

The **Preludio metadata management system** supports editing, managing and searching for meta

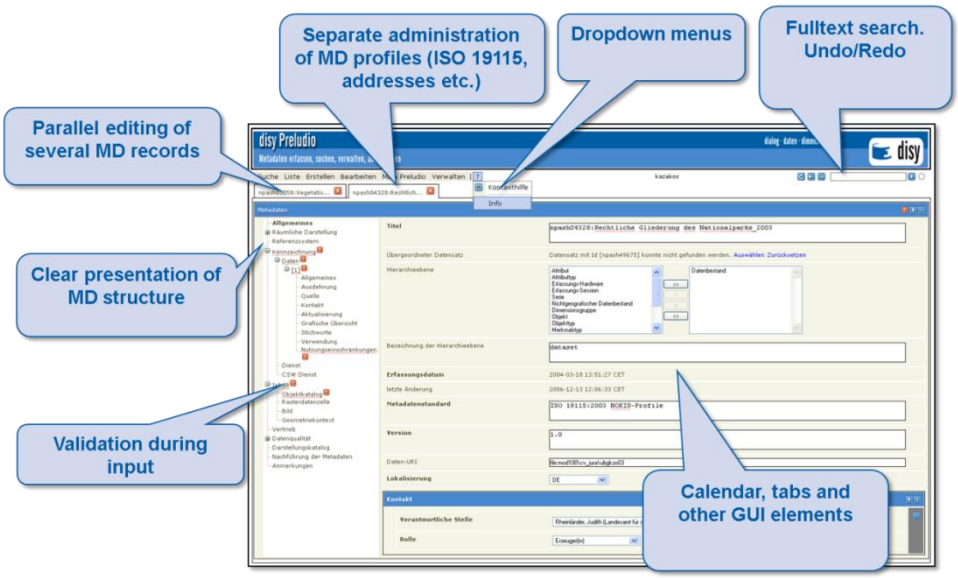

Figure 5: Some Preludio GUI Design Principles and Functionalities

data that describe data and services in a spatial data infrastructure. Preludio is compliant with INSPIRE, OGC and ISO standards 19115/19119/19139. It implements a metadata server following the Catalog Web Services standard CSW 2.0.2. ISO API 1.0. Preludio has a Web application similar to the MS Windows GUI principles, with a full-text and extended search, as well as an integrated map viewer. A main design goal for Preludio was user-friendliness and usability, putting much emphasis on a clear presentation of the metadata structure. It allows parallel editing of several metadata entries and provides manifold mechanisms for automated validation during the input. Such mechanisms comprise specification of default values for metadata fields, default values depending from the actual input, automated validation against schema and INSPIRE implementation rules, as well as calling the German GDI-DE INSPIRE test suite. The metadata is stored in a relational database. Both the metadata schema and the system GUI can be freely configured. One application can simultaneously support several metadata profiles. Preludio allows the integration of external Web Services. Preludio is widely used in German environmental agencies.

First integration experiences will be available during late 2014. A full integration with Preludio is planned for mid of 2015. Conceptually, one of the more promising upcoming challenges is an understandable, yet powerful GUI design and interaction design for the visualization tool working with the **SemanticExplorativeSearch** module.

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

References

[1] Albertoni R., De Martino M., Di Franco S., De Santis V., and Plini P., "EARTh: An Environmental Application Reference Thesaurus in the Linked Open Data Cloud," *Semantic Web*, 2014.

[2] Albertoni R., Gómez-Pérez A.: Assessing linkset quality for complementing third-party datasets. EDBT/ICDT Workshops 2013: 52-59.

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

[4] Heath T. and Bizer C., *Linked Data: Evolving the Web into a Global Data Space (1st edition).* Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, pp. 1-136. Morgan & Claypool, 2011.

[5] De Martino, M. and Albertoni, R., "A Multilingual/Multicultural Semantic-based Approach to Improve Data Sharing in a SDI for Nature Conservation", *International Journal of Spatial Data Infrastructures Research*, vol.6, ISSN 1725-0463, pp. 206-233, 2011.