Semantics for the „Long Term Ecological Researchers“

Herbert Schentz, Johannes Peterseil & Michael Mirtl,
Umweltbundesamt GmbH (Environment Agency Austria)
Spittelauer Lände 5, 1090 Vienna, Austria

Abstract

In the following paper the application of semantics for LTER-Europe, a community of researchers dealing with long term ecosystem research, is described. The needs from this community can be seen as representative for the environmental domain within research and public administration. Within the ALTER-Net project a test for semantic data integration has been carried out, where dislocated, very heterogeneous data were mapped to a common ontology, thus allowing a seamless and homogenous data integration. This test showed, that it is feasible, but a lot of issues have to be overcome. One lesson, learned out of this test was, that one comprehensive, complete conceptual model for this big domain cannot be established in a reasonable timeframe. The work has to be split up into several steps, must make use of work already done and a simple start is needed.

We developed a thesaurus as a simple start on semantics and interlinked it with other existing vocabularies, which are important for the community (GEMET, EUROVOC, AGROVOC). EnvThes aims to cover the concepts for environmental monitoring and experimentation. So far this thesaurus is used for controlled keywords within the metadata system DEIMS and in the future the concepts should be mapped to an ontology and data should be annotated with the concepts. Building an ontology, derived from ISO19156 (observation and measurement) and annotate the underlying data would be the next steps.

1 Scope of the “Long Term Ecological research”

The Long-Term Ecosystem Research (LTER) is an essential component of the world-wide efforts to better understand ecosystems, their functioning and the effects of driving factors on its processes. LTER Europe (Long Term Ecological Research Network[1]) is the European contribution to this effort. It is a network comprising around 455 long term observation sites (420 LTER sites and 35 LTSER Platforms) organised in 22 national LTER networks across Europe covering a broad biogeographic gradient as well a important ecosystem types. The research topics cover the understanding of structure, functions of ecosystems, and their response to environmental, societal and economic drivers. LTER contributes to the knowledge base informing policy and to the development of management options in response to the Grand Challenges under Global Change.

From the beginning (around 2003) the design of LTER-Europe has focussed on the integration of natural sciences and ecosystem research approaches, including the human dimension. LTER-Europe was heavily involved in conceptualizing socio-ecological research (LTSER). As well as LTER Sites, LTER-Europe features LTSER Platforms, acting as test infrastructures for a new generation of ecosystem research across European environmental and socio-economic gradients.

LTER Europe is organized along Expert Panels[2] providing best practice examples and a forum for the further discussion. With the submission to the ESFRI roadmap in 2015 effort was made in order to establish LTER as European scale Research Infrastructure (RI).

The provision and sharing of information (metadata and data) is one of the core components of the efforts taken. In LTER Europe data are managed and quality controlled by the different organisations running the LTER sites and data are provided on a national or local scale, following standardized protocols (e.g. UNECE ICP Integrated Monitoring, UNECE ICP Forest, TERENO, etc.) and methodologies. To streamline the process, LTER Europe develops (based on European scale projects)

[1] see , http://www.lter-europe.net/
[2] see http://www.lter-europe.net/ep-tf

tools and strategies to publish, disseminate and leverage the use of existing data trying to enforce a more open data policy. The resulting architecture aims to provide a scalable solution for data management within the LTER network. The concept of a 'data node' combining a series of functionalities (e.g. metadata provision, data storage, data services) is introduced in order to provide nodes for a distributed data infrastructure. This data nodes can either exist, e.g. TERENO Data Portal[3] or are created within the eLTER H2020 project, which supports this activities. In this understanding, different data provider / nodes can also be registered in other networks, e.g. European or national thematic data portals using standard services from information exchange. Information from the different data nodes is harvested and provided by a central discovery portal, the 'Data Integration Platform'. This will provide the central point to link LTER data into other networks (e.g. DataOne, EUDAT, GEOSS, etc.) and to provide a single point of access for the research community as well as for end users.

Derived from the scope and organizational structure of LTER Europe, the data can be characterized as follows:

- LTER covers broad range of scientific domains and ecosystem types
- LTER involves a large user community of players, ranging from data providers to data consumers
- data policies are varying between the data providers
- a high heterogeneity with respect to data formats, data management practices and solutions
- a high heterogeneity with respect to the implementation and use of semantics for data description

In order to provide an integration of data across the domains and institutions we have to face these challenges. The development of common semantics and the integration into the workflow of data acquisition, management and analysis is one of the crucial aspects in the environmental domain.

2 Solutions

A common semantic is the backbone for the integration, use and interpretation of data coming from different distributed data providers. In order to cope with this, LTER aimed to define or adopt structures for data documentation (e.g. EML, ISO19115) and exchange (e.g. ISO19156), as well as common semantics, which allow the harmonized, seamless description and presentation of the data. For LTER Europe this was done in a series of European scale projects aiming for the integration of data providers in the domain.

2.1 SERONTO – a concept for semantic data integration

The Network of Excellence project ALTER-Net[4] started in 2004 developing inter alia a concept for semantic integration of ecological data. The work followed the concept of semantic annotation, mapping heterogeneous and distributed data to a common conceptual model and thus integrate the data seamlessly.

This was implemented in three consecutive steps: a) design of a common test environment, b) the development of the common core ontology SERONTO[5], and c) finally a proof of concept by testing the mapping to the Ontology (OntoBroker).

Despite the fact of a successful test of the proof of concept, the work was complicated by several drawbacks. First of all, the development of the common core ontology SERONTO, which aimed to cover all the domains within the LTER community, took too much time. This resulted in a lack of time for deriving domain specific ontologies from the core ontology. Secondly, although the application of the core ontology for mapping existing data to it within OntoBroker worked nicely, at that time no software for further processing those data existed (at that time OntoBroker neither had an LD interface

[3] see http://teodoor.icg.kfa-juelich.de/overview-de
[4] see http://www.alter-net.info/about-alter-net#sthash.qx2owSo8.dpuf
[5] see https://www.umweltbundesamt.at/fileadmin/site/daten/Ontologien/SERONTO/SERONTOCore20080812.owl

nor a SPARQL endpoint). And finally, many data provider within the ALTER-Net community did not even have a database, the precondition for applying a service.

2.2 EnvThes - a domain- specific and interlinked thesaurus

One of the lessons learned out of the establishment of SERONTO, was, that it is impossible to create "a single conceptual framework" for the integration of data across the LTER community within a manageable timeframe. Therefore we looked for an alternative approach, which allows for a) starting with a simple structure for the concepts (SKOS / RDF), b) to work on small portions, that are clearly laid out, c) to include already existing concepts and the establishment of links to the original concept definition, and d) avoiding complicated guidance

This was realized by the development of the interlinked SKOS/RDF thesaurus EnvThes[6], which was developed within the Life+ project EnvEurope[7] and ExpeER[8] (FP7). EnvThes aimed to cover the concepts needed to describe data from the long term ecological domain, in order to firstly provide keywords for dataset metadata and secondly in the long term allow for semantic annotation of data itself. EnvThes was developed in a collaborative manner.

2.3 Development steps

At the beginning of the work on EnvThes a screening of existing and relevant controlled vocabularies was undertaken. The US-LTER controlled vocabulary[9] was chosen as backbone for EnvThes, extended by concepts needed within the projects EnvEurope and ExpeER, refined by Definitions. To provide. The need to include the EUNIS habitats and the INSPIRE spatial data themes arose out of European obligations and the extension with concepts of the AnaEE thesaurus was necessary for the description of experimental data. In a first step taxa of the Catalogue of Life have also been included, but was dropped later on, as the CoL team intends to expose their list as LD.

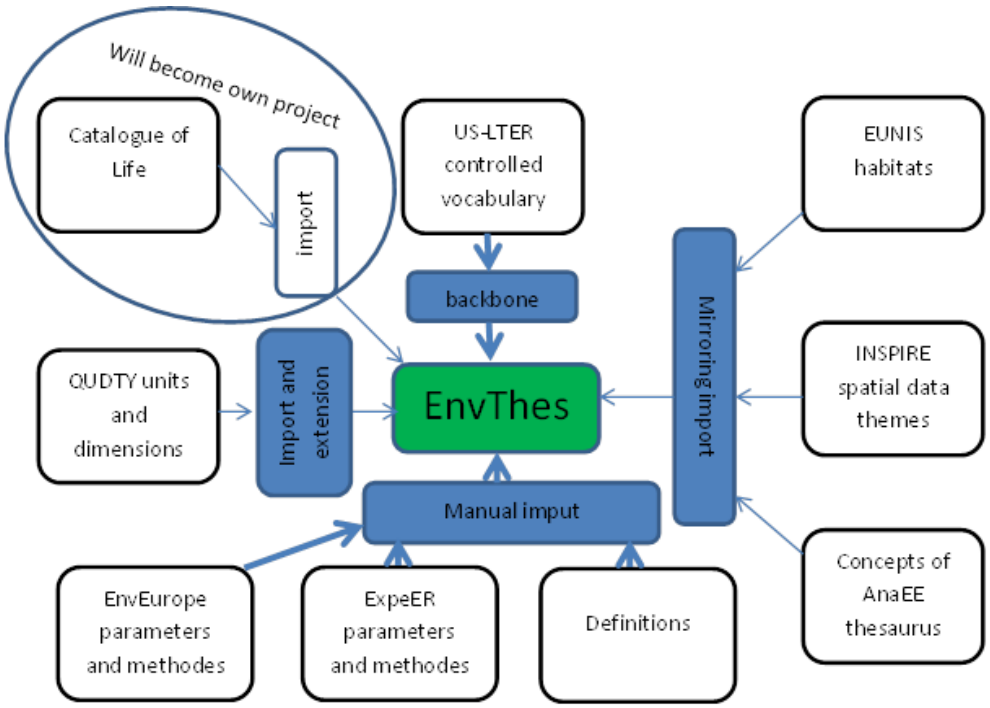

Fig. 1 *Overview on the existing vocabularies taken into account for EnvThes*

[6] see http://vocabs.lter-europe.net/EnvThesDev.html
[7] see http://www.enveurope.eu/
[8] see http://www.expeeronline.eu/
[9] see http://vocab.lternet.edu/vocab/vocab/index.php

New concepts were defined by contributors during a reviewing process in common EnvThes development meetings. For new concepts a clear definition was needed as well a check if the term not already existed.

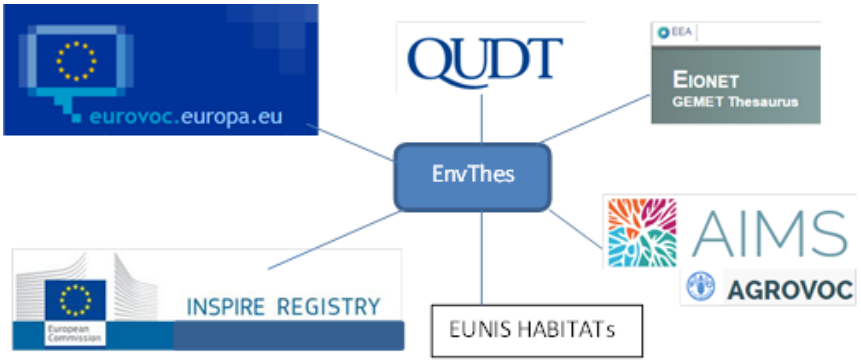

Fig. 2 Overview on vocabularies linked to EnvThes

As second step links to other vocabularies, which are relevant for the work of LTER-Europe partners have been established (see Fig.2): EUROVOC, as "official vocabulary" of the European commission, GEMET, the vocabulary of the European Environment Agency, INSPIRE spatial data themes as they are needed for compliance with the INSPIRE directive , AGROVOC the FAO vocabulary, a comprehensive standard within the agricultural domain, QUDT an ontology for units and dimensions, first introduced by NASA. These linkages are important to avoid "semantic silos".

2.4 From thesaurus to ontology ... (observation and measurements – extended)

In the final review of the main structure, it turned out, that a configuration beyond the simple Elements of SKOS would be desirable. Currently we are in the phase that well agreed ontologies within the domain of ecology are arising and a lot of "best practices" are established. One of these ontologies is ISO 19156, "Observation and measurement". This Ontology has been taken into account (although just in the background), when considering and developing the root elements of EnvThes, thus making it possible to further establish the link between the controlled vocabulary and the ontology. In a further phase the concepts of EnvThes might become Individuals of an ontology derived from ISO 19156.

3 Open Issues

Despite the current use of EnvThes as source for metadata keywords many other applications within the LTER Europe data infrastructure are eligible. In the following the most important issues are listed which will be addressed in the near future.

- Conceptual work on Vocabularies

As already pointed out above, it turned out, that the establishment of an ontology as extension of EnvThes would be highly desirable. This means, that we have to take an existing ontology (most probably ISO 19156, O&M in OWL) and extend it by specifically needed classes. Furthermore, the assignment of concepts of EnvThes to those ontology classes is needed. This could be done most probably by adding them as Individuals of the respective classes in the ontology. This for sure needs resources and flexible tools allowing for a collaborative work in the development of the ontology and the integration of the concepts of EnvThes.

- Semantic annotation of data

One of the key use cases within the LTER data lifecycle is linking the data contents with the controlled vocabulary applying semantic annotation. This should be done as early as possible by the person leading the data acquisition and knowing all the details on the data. This information has to be

described in a human and machine readable way in order to allow for a correct reuse and interpretation of data. Despite the need for this information the efforts and resources needed are considerable and often are in direct conflict with other tasks within data acquisition and maintenance.

Therefore usability of software and profound support for the scientist and/or technician is crucial for the success. This applies for a) the easy and standardized access to the relevant vocabularies, b) the availability of predefined "default" descriptions (small default ontologies) for certain domains of data, c) a supporting processes helping to find controlled vocabularies on the base of the existing measurement outcomes, and d) the possibilities to add unstructured information (e.g. images, movies, sketches, notes) to the data in order to ease the interpretation and use.

- Provenance and provenance tracking

The correct interpretation and reuse of data is strongly dependent on the possibility to understand the processes, which have been applied to generate the data. This includes information e.g. on the data acquisition processes, the applied QA/QC procedures, the data aggregation methods or data merge or information on models applied to generate the data. The W3C provenance working group elaborated an extensible ontology for that purpose (prov-O) , which we already considered to use.

- Semantic web services

Within the EnvEurope project we have tested D2RQ to expose data of a relational database as linked data services and via a SPARQL endpoint. This test showed, that the task can principally be done, but that there are still some issues to overcome. One crucial issue is the need of a cache for the triples. Converting the relational structure to Triples on the fly is nice for smaller databases but has its limits of tables with about 100.000 records.

- Use of semantics for "drill into"

The links, which have been established between "data" – the results of observations and measurements and concepts, have to be resolvable later on. This means that all representations of the data and their evaluations (Tables, GIS representations, outcomes of statistical processes, results of model calculations, …) have to offer the possibilities to follow established links or links derived from input data, to allow for "drilling into" all the information , needed for a correct interpretation of the data.

4 Outlook

We are convinced, that semantic data integration can help to reuse data, thus capitalizing money, invested in observations and measurements, but that there are still a lot of barriers that have to be overcome. Controlled vocabularies are certainly cornerstones in the process of semantic data integration and interlinking different existing vocabularies is an important and not so difficult process on the way to bridge differing existing semantic worlds.

An important question to be addressed by the community is 'Why aren't we further on the way of semantic data integration?' On the one hand, of course, because some bits and pieces of software are still missing. On the other hand, because a lot of "good practices" are still contradicting – hopefully the cooperation between OGC and W3C will help to bridge one set of contradictions. But last and possibly most important, because the scepticism of scientists and/or data producer towards opening of "their" data is still big. They have differing readiness to do so, different licensing opinions.

With the work on SERONTO and the development of EnvThes for LTER Europe the way into semantic annotation and integration was laid. Being used currently a plain provider of keywords with the future work a stronger integration into the workflows and data provision is planned. A first step will be the implementation of a life link between DEIMS, the metadata repository for LTER Europe, and EnvThes.