# Integration of semantics, data and geospatial information for LTER


## Abstract

The long term ecological monitoring and research network (LTER) in Europe[1] provides a vast amount of data with regard to drivers and impacts of environmental change (Mirtl & Krauze 2007, Mirtl et al. 2013). In a test within the EnvEurope (LIFE08 ENV/IT/000399[4]) and ExpeER projects those data have been exposed via a linkedData service and SPARQL endpoint (D2RQ) and an OGC SOS to find out best practices for data services. A SNORQL editor, a SPARQL to EXCEL tool and, most important of all, an R SPARQL plugin have been tested as clients. This access worked quite well and SPARQLing the D2RQ service and the thesaurus SPARQL endpoint proofed feasible.

When using this architecture for accessing distributed services, however, query broker and/or caches are recommendable for performance reasons.

Top requirement, however would be the integration of spatial data and semantics. Although geoSPARQL is already a W3C and OGC recommendation and a geoSPARQL extension to the D2RQ service exists, we could not test that so far because of the lack of geoSPARQL clients.

SPARQLing data sources is nice for the IT-trained domain specialists, our “normal” domain specialist, however would need applications with simple and easy to use GUIS, based on the SPARQL interface.

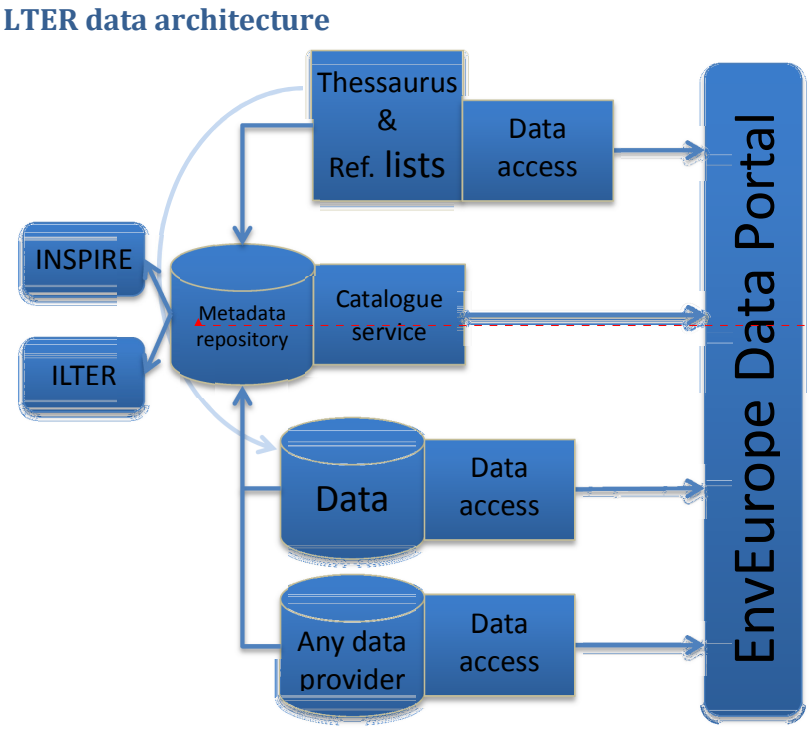

## LTER data challenges

The long term ecological monitoring and research network (LTER) in Europe[1] provides a vast amount of data with regard to drivers and impacts of environmental change (Mirtl & Krauze 2007, Mirtl et al. 2013). A variety of ecosystems are monitored touching research topics of several different disciplines, like biodiversity research, soil science, air pollution measurement, groundwater and surface water analysis, meteorology, hydrology and others. In addition to the variety of disciplines involved, data management and analysis is fragmented and carried out by several dislocated institutes which are responsible for the monitored sites. Because of the distributed responsibility, the implemented data management systems vary from collections of simple spreadsheet tables, via distributed domain specific databases to centralized data management systems. Within those systems data are organized using different data models, and related vocabularies, taxonomies and code lists. Existing semantic de facto standards like the EUNIS habitat list[2], the Catalogue of Life, or the World Reference Base for Soil Resources[3] are often not considered depending on the autonomous decision of each data manager and provider.

Overcoming the described semantic heterogeneity, resulting from the variety of disciplines and institutes, is crucial for the intended exchange of metadata and data. The establishment and use of a common set of exactly defined concepts is a first step into the direction of semantic harmonization; using those concepts to annotate data and metadata and finally using them

directly in the local data management systems, should be the subsequent activities (Schentz et al. 2005, Adamescu et al. 2010).

LTER Europe and related projects like EnvEurope (LIFE08 ENV/IT/000399[4]) or ExpeER[5] are working towards a harmonisation of methods and data flows for long term ecological research, monitoring and experiments.




Formatiert: Schriftart: 10 Pt.

Fig 1. LTER data architecture (© Johannes Peterseil)

A DRUPAL based **metadata container** is provided to manage data related and site related metadata. Those metadata are tagged with keywords from the common thesaurus.

The metadata repository provides a download service in EML format and INSPIRE (close to ISO19115) format and a CSW (OGC- catalogue service) interface.

So far there also is a central **data cache**, used for experiments with data access services:In order to provide recommendations for the LTER Europe network a two-fold strategy was applied.

First a linkedData service using D2RQ was established providing online access to data. With a SPARQL plugin for R statistical software the data from distributed resources can directly be used in the analysis or displayed. Second a SOS server and SWE client provided by 52°North are used to evaluate XML based data services. These services can also be directly consumed by R statistical software or other clients.

A SKOS/RDF based thesaurus (EnvThes) serves as a source of keywords for the metadata system **and as knowledge base,** to which pointers (URIs of the concepts), entered into the data records, are pointing.

## LTER Linked Data service (proof of concept)

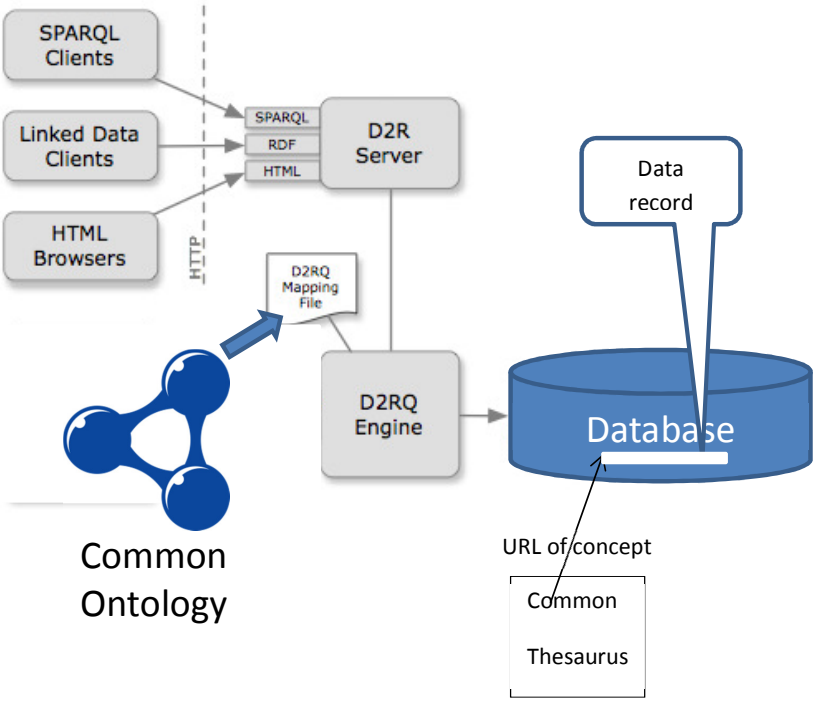



Fig 2: D2RQ service (with regard to the mapping to a common ontology)

It should be proven, that exposing data of a relational database via a linkedData interface and SPARQL endpoint, is possible and that existing knowledge bases can be integrated. Furthermore the question also was, whether or not this service could be used to link distributed databases.

Because of the lack of time, the default ontology generated by the D2RQ software, was used as common ontology for the mapping process. With some little modifications it served as bases for the class-mapping and EnvThes, the thesaurus started within the EnvEurope project, was used to map some database records to a common vocabulary (instance mapping).

As SPARQL clients we tested:

(1) The built- in SNORQL editor
(2) A download service for EXCEL spreadsheets
http://www.snee.com/sparql/spreadsheetSPARQL.html
(3) A SPARQL plugin for the statistic package R
http://linkedscience.org/tools/sparql-package-for-r/

All the 3 clients proofed to work sufficiently, offering an IT-trained domain expert ( at an IT level that could be described as “can work with MS-access”), the possibility to start further analysis of data starting with the SPARQL query.

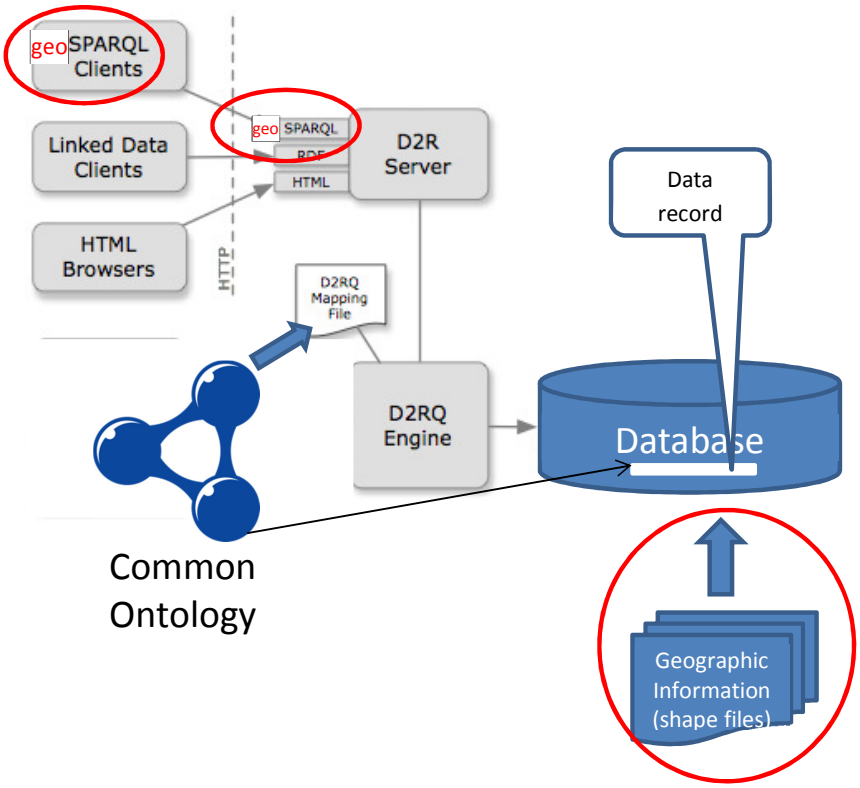
When using the R plugin it is additionally possible to establish some predefined queries for reuse by people who are not familiar with SPARQL.

Within our tests we also could do a SPARQL query over the D2RQ service and EnvThes, thus eg. getting a list of parameters from the D2RQ service and their definitions from EnvThes.

The performance of the D2RQ service is not exhilarating and the service does not do any query optimization. But the general test- outcome was, that this service offers the possibility to map data to a common vocabulary and make several of such services accessible for distributed SPARQL queries.

## Further work needed

(1) Top priority of future work is the full integration of spatial information and semantic information. At the Austrian environment agency, the integration of geographic data




and other information within ORACLE databases is state of the art. Mostly geographic information results from import of shape files.
This geographic information usually is displayed with ESRI tools working in a local client server environment or exposed via WFS (OGC web feature service) or WMS (OGC web map services).
The possibility to offer the geographic data via a geoSPARQL endpoint technically is ready, but never was tested by us, because of the lack of geoSPARQL clients.

(2) Whilst SPARQLing data is a nice possibility for domain experts with advanced IT skills, we are so far missing tools and/or toolsets for the not so IT-trained users.
(3) Performance is crucial for domain specialists who want to do evaluations based on distributed data and/or knowledge. To our experience querybrokers and caches at the site of the analyst are strongly needed. For us real-time analysis is not dominant, so that a harvesting mechanism bringing regularly data to the data mining site could be a good solution.