Exposing the SeaDataNet metadata catalogues via SPARQL endpoints

Chris Wood, British Oceanographic Data Centre (UK), c.c.wood@gmail.com
Alexandra Kokkinaki, British Oceanographic Data Centre (UK), alexk@bodc.ac.uk
Adam Leadbetter, Marine Institute (Ireland), adam.leadbetter@marine.ie
Rob Thomas, Marine Institute (Ireland), rob.thomas@marine.ie

Many scientific disciplines have metadata catalogues managed by a designated responsible organisation. The content of these catalogues can vary by discipline, but may contain information about institutions who contribute data to central data repositories, projects that have been carried out within a particular field, or individual datasets. Within the European oceanographic science community, metadata catalogues have been managed under the SeaDataNet infrastructure. The five catalogues (EDMED: for datasets, EDMO: for organisations, EDMERP: for projects, EDIOS: for observing systems, and CSR: to describe cruise summary reports) have long been available online through individual web based search interfaces via the organisation that hosts the catalogue. However, such interfaces implicitly limit how the queries can be conducted, and how results can be viewed. Such limitations can be removed through the development of Application Programming Interfaces (APIs).

We have led the API development for all five catalogues, and have taken a Linked Data approach to the implementation. This approach requires the catalogues to be stored in triplestores, a form of graph database, with querying available using SPARQL, a query language for triplestores analogous to SQL for relational databases. The query interface is publicly available over HTTP, allowing the whole catalogue to be openly queryable. The true strength of the triplestore approach is the ability to conduct federated queries across the different SPARQL endpoints which simulates a join between tables in a traditional table-based database, but without the need for subsets of the data to be located within the same database. A sixth triplestore, supporting the NERC Vocabulary Server (NVS), underpins the content of the catalogues by providing lists of standardised terms that provide consistency and semantic harmonization across the catalogues. Federated queries can be used to determine the theme of a dataset, described in NVS terms, by simultaneously querying the SeaDataNet dataset triplestore and the NVS triplestore.

In this presentation, we will show the advantages that can be gained by both the catalogue publisher and the end user, the steps needed to setup both a new triplestore and the corresponding SPARQL endpoint and the ease by which this software stack can be implemented, as well as the options available to the publisher. We will finish with the lessons learnt in this project, and the future work that will be carried out to further enhance the service to the users of the SeaDataNet infrastructure.