Development of Linked Data Services to Support Widespread Exposure of Data

Chris Wood, British Oceanographic Data Centre (United Kingdom), chwood@bodc.ac.uk

Historically, environmental data centres have tended to store all the metadata and data they control on internal databases. The lack of public interfaces to such databases means that the dissemination of the stored data is then a relatively labour intensive process, whereby a scientist or other interested user contacts the data centre requesting a particular piece of data, a data centre employee retrieves the relevant data, and emails it to the user. In the last few years, web technologies have allowed fairly complex websites to be developed which allows more of the metadata to be exposed and for increasing amounts of data and metadata to be downloadable directly by users. However, this approach still presents limitations to both the data centre and end users; data centres have to ensure that the development of such systems provides an intuitive interface for end-users to interact with – an unintuitive system means that end-users are likely to engage less with the interface and revert to contacting the data centre directly. The data centre also must ensure such a system is up to date with trends in interface design and searching ability to keep up with the experience that ever-increasingly technically-savvy web users expect. End-users, meanwhile, have to become familiar with the search interface for every service they wish to interact with.

This particular paradigm is being solved by some data centres, including the British Oceanographic Data Centre (BODC), by using an emerging World Wide Web Consortium (W3C) standard: the Resource Description Framework (RDF) Model. RDF is a data structure where the description of the data, as well as the relationships between both data and metadata is named. These descriptions and relationships are made up of three elements, and, as such are normally referred to as triples. Triples can be stored in specialist databases, known as Triplestores, and often allow the triples to be queried using the SPARQL query language. The relationships between data and its metadata is often described using Unique Resource Identifiers (URIs), and as such are ideal for exposing data on the web. Where a triplestore provides both a web-accessible front-end and the ability to query with SPARQL (commonly known as a SPARQL endpoint), it provides a convenient method for mass amounts of data to be exposed with little effort needed by a data centre. Triplestores become particularly useful when the data they hold is richly described using published ontologies (both very generic and subject-specific), and thus allows end-users to search the data using terms they are already familiar with.

Up until now, exposure of data via triplestores has been slow due to a chicken-and-egg style problem: data centres have been reluctant to spend time and effort implementing triple stores while there are few subject-specific ontologies that will allow their data to be searched effectively; meanwhile, ontology authors have not been creating subject-specific ontologies when there are too few triplestores that would use the ontology. This has led to a slow uptake of the technology by data-centres, with the number of end-users who are experienced in the use of the technology also trailing behind. However, the technology seems to be reaching a critical mass: as the number of data centres who are interested in using triplestores increase, the number of suitable ontologies increase, and the number of users who are experienced in the use of them, and expect data centres to provide them as a means to expose their data, increase. This in turn leads to more data centres providing SPARQL endpoints, and encourages ontology authors to provide richer and higher-quality ontologies.

This presentation will show how the BODC has implemented SPARQL endpoints for two distinct use-cases over the past year. First, I will show how BODC have exposed the vast majority of the metadata describing data series that we hold via an implementation of JENA, an open-source triplestore with an associated API that we have used to ensure that the triples that we store and expose are an accurate

representation of the metadata held in the underlying database. As this triplestore is kept in sync with our publicly-available data, the SPARQL endpoint provides a convenient method for end-users to query the data and metadata that we hold. I will explain how the wide range of ontologies (both generic, such as Dublin Core (dc) and Simple Knowledge Organisation System (skos), and subject specific, such as the Ocean Data Ontology (odo) and the Marine Metadata Interoperability Semantic Web Services (MMISW)) that we have chosen to use to describe the data are a useful aid into discovering our data, particularly by users who already have the technical knowledge in using a SPARQL endpoint. I will also explain how we plan to expand the current service by implementing machine-to-machine data exchange. This will enable data to be automatically transferred or downloaded, in a range of file formats due to content-negotiation or URL structure that we will implement, as a result of a single, simple SPARQL query. We hope that such a system will transform the user experience associated with the search and download of specific datasets, either by using a native SPARQL endpoint, or subsequently by using simple user interfaces that have the potential to be built on top of the simple HTTP-based API that such an endpoint provides.

Secondly, I will show how a SPARQL endpoint has been used as the API for a particular application – in this case to support the development of a metadata portal that has been used to aid implementation of the Marine Strategy Framework Directive in the Celtic Seas Region. This has been used to support policy makers, special-interest groups, users of the marine environment, and other interested stakeholders in the legislation, which mandates 'Good Environmental Status' to be maintained or achieved by 2020, through a series of Programme of Measures. The metadata portal has been built to provide a signposting service to datasets that are relevant to MSFD within the Celtic Seas. Although the metadata are stored in a traditional RDBMS, they are exposed as linked data via the D2RQ platform, allowing virtual RDF graphs to be generated. This also ensures that the metadata has the widest possible reach, and allows us to expose the service, and associated metadata, as an API. SPARQL queries can be executed against the published end-point allowing any user to search the metadata.

As with the triplestore which is used to expose BODC's data series, we have again mapped a wide range of relevant ontologies to the metadata. However, in this case, we have used D2RQ's mapping language, based on the turtle format, to generate the mappings. The ontologies used (e.g. The Provenance Ontology (prov-o), Ocean Data Ontology (odo), Dublin Core Elements and Terms (dc & dcterms), Friend of a Friend (foaf), and Geospatial ontologies (geo)) allow users to browse the metadata, either via SPARQL queries or by using D2RQ's HTML interface. In this instance, the metadata were further enhanced by mapping relevant parameters to the NERC Vocabulary Server, itself built on a SPARQL endpoint, therefore allowing federated queries to be written.

Additionally, we have built a custom web front-end to enable users to browse the metadata and express queries through an intuitive graphical user interface that requires no prior knowledge of SPARQL, and therefore enables users to effectively filter, sort, or search the metadata. As well as providing means to browse the data via MSFD-related parameters (Descriptor, Criteria, and Indicator), the metadata records include the dataset's country of origin, the list of organisations involved in the management of the data, and links to any relevant INSPIRE-compliant services relating to the dataset. As the MSFD timeline requires Member States to review their progress on achieving or maintaining GES every six years, the timely development of this metadata portal will not only aid interested stakeholders in understanding how member states are meeting their targets, but will also show how a linked data approach, along with the associated technologies, can be used effectively to support policy makers, scientists, and other end-users.