Cataloguing Ocean Data at Web Scale

Adam Leadbetter, Marine Institute (Ireland), adam.leadbetter@marine.ie
Andrew Conway, Marine Institute (Ireland), andrew.conway@marine.ie
Tara Keena, Marine Institute (Ireland), tara.keena@marine.ie
Will Meaney, Marine Institute (Ireland), will.meaney@marine.ie

In order to allow users of marine data to discover (or "F"ind, Wilkinson et al., 2016) the datasets which meet their needs, a data cataloguing solution is required (Leadbetter, et al., 2020). Data catalogues should then point users to accessible data download services and where possible the catalogues should use common vocabularies for interoperability and best practice metadata profiles in defined standards. Data cataloguing may take place at an institutional or national level, and for needs such as marine spatial planning or reporting under the Marine Strategy Framework Directive. The datasets described by the local catalogues may be then aggregated to a regional or an international scale.

As such, Data Catalogues are often required to feed into other, aggregated Data Catalogues. For this to be achieved, the base metadata schema of the Data Catalogue system must be mapped in a crosswalk to the metadata schema of the target Data Catalogue. This will provide syntactic interoperability and, if controlled vocabularies are used to populate the fields of the source Data Catalogue and mapped to the controlled vocabularies of the target metadata schema, semantic interoperability (Schaap and Lowry, 2010). Where the aggregation takes place in a more generic data portal, such as a data.gov portal, or uses a generic metadata profile, such as Schema.org, the crosswalk may lead to some loss of detail compared with the source metadata.

In the past, metadata validation, crosswalks and other related tasks have been achieved either through one-off calls to scripts or services or through batch runs of jobs on a schedule. However, applying principles from modern software engineering approaches to these data management tasks yields an alternative, web-scale approach to both metadata publishing and metadata engineering tasks.

Continuous Integration (CI) is a process by which a team of software developers contribute changes to a single working copy of a code base. CI is reliant on a source control system being used to manage the code base. Once code is committed to a CI pipeline, Continuous Delivery tools are used to build the software product into a deployable artifact. If the build fails, the developers will receive warnings as to why but the previous, fully built version of the software artifact will remain available. Once a build is completed, Continuous Deployment tools can seamlessly push the new software artifact to users.

The Marine Institute, Ireland has undertaken a pilot project applying this paradigm to data managers who are responsible for metadata cataloguing. The data managers are given access to a source control repository, through the GitHub platform. A data manager may commit a completed metadata record in ISO19139 XML format to a folder in the source control repository. Once the new metadata record, or records, have been pushed to the source control repository, a number of Continuous Delivery tasks are automatically started through the TravisCI platform. These tasks include: generating DataCite metadata kernels for use in minting digital object identifiers for data citation; producing HTML landing pages with Schema.org annotations (Leadbetter, et al., 2018) to allow for indexing in Google's Dataset Search; and creating Global Biodiversity Information Facility records. Continuous Deployment is achieved through the TravisCI tasks commuting back to the source code

repository and general access to HTML representations of the metadata through web hosting via GitHub Pages.

The Continuous Delivery tasks have been scripted in Python, and are based around a class which has been developed to provide access to the information held within a metadata record. The various target export formats are templates within a templating framework, such as Jinja. Through Python code, the templates have access to data from instances of the metadata class.

Whilst the approach described in this paper has been undertaken as a proof-of-concept, it is anticipated that it will become operational over time and the approach could be used for future aggregations such as the Intergovernmental Oceanographic Commission's Ocean Data and Information System (http://odis.iode.org/) proposed under the UN Decade of the Ocean.

References

Leadbetter, A., Thomas, R., Shepherd, A., Fils, D. And O'Brien, K. (2018). The place of Schema.org in Linked Ocean Data.

Leadbetter, A., Meaney, W., Tray, E., Conway, A., Flynn, S., Keena, T. Kelly, C. and Thomas, R. (2020). A modular approach to cataloguing marine science data. Earth Science Informatics. https://doi.org/10.1007/s12145-020-00445-w

Schaap, D. and Lowry, R. (2010). SeaDataNet - Pan-European infrastructure for marine and ocean data management: Unified access to distributed data sets. International Journal of Digital Earth 3(Sup. 1): 50-69.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18