



An ODIP effort to map R2R ocean data terms to international vocabularies

Renata Ferreira (1), Karen Stocks (1), and Robert Arko (2)

(1) Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, United States (rferreir@ucsd.edu),

(2) Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York, United States

The heterogeneity of terminology used in describing data creates a barrier to the efficient discovery and re-use of data, particularly across institutional, programmatic, and disciplinary boundaries. Here we explore the outcomes of a student project to crosswalk terms between the Rolling Deck to Repository (R2R) program and other international systems, as part of the Ocean Data Interoperability Platform (ODIP).

R2R is a US program developing and implementing an information management system to preserve and provide access to routine underway data collected by U.S academic research vessels. R2R participates in ODIP, an international forum for improving the interoperability and effective sharing of marine data resources through technical workshops and joint prototypes. The vocabulary mapping effort lays a foundation for future ocean data portals through which users search and access international ocean data using familiar terms.

R2R describes its data with a suite of controlled vocabularies (<http://www.rvdata.us/voc>) some of which were developed locally or are specific to the US. The goal of this student project is to crosswalk local/national vocabularies to authoritative international vocabularies, where they exist, or to vocabularies widely used by ODIP partners. Specifically, R2R developed the following crosswalks: R2R science party names to ORCID person identifiers, UNOLS ports to SeaDataNet Ports Gazetteer, R2R Device Models to NVS SeaVoX Device Catalog, and R2R Organizations to the European Directory of Marine Organizations (EDMO). Mappings were done in simple spreadsheets using synonymy relationships only, and will be published as part of the R2R Linked Data resources.

The level of success in crosswalking was variable. The majority of ports were successfully mapped. Differences in the character sets (i.e. whether diacritic marks were used) caused automated matching to fail occasionally, but the number of ports was small enough that these could be manually reviewed. Both organizations and device models have initial mappings, and R2R will propose new terms to the EDMO and SeaVoX Device vocabularies to complete coverage. Mapping to ORCID identifiers was abandoned (though R2R will still hold and expose them when supplied by the data provider). Most ORCID entries do not contain insufficient metadata to confirm potential mappings: the match of a family and given name was considered inconclusive without further support. ORCID also does not assign identifiers posthumously, which is occasionally necessary for historical data in R2R.