Chapter 4
Loose Integration of Local Information to Generate Collaborative Marine Science Knowledge

Adam Leadbetter
British Oceanographic Data Centre, NERC, UK

Cynthia Chandler
Woods Hole Oceanographic Institution, USA

Robert Arko
Columbia University, USA

Adam Shepherd
Woods Hole Oceanographic Institution, USA

Roy Lowry
British Oceanographic Data Centre, NERC, UK

ABSTRACT

This chapter focuses on improved access to marine science data, enabling researchers to generate new information and knowledge products. The history of controlled vocabulary developments in marine sciences, from paper publications to the Semantic Web, is explored in detail. This history is being furthered through the publication of Linked Open Data, meaning: the publication of clearly identifiable entities; a simple, universal mechanism for retrieving resources; a generic graph-based data model; and publishing explicit relationships to other resources. Progress towards Linked Open Data for marine science is reported in this chapter. As shown by the Data-Information-Knowledge ecosystem, the approach of "small pieces of data loosely joined" provides presentation and organisation to data, which creates information. The use of query endpoints to integrate this information from multiple locations into a knowledge base, which required active collaboration between cooperative partners to truly generate new knowledge and to address emerging science questions, is described.

INTRODUCTION

Data collected during the course of oceanographic research expeditions have traditionally been archived close to the data originator: either in institutional or project based Data Assembly Centres (DACs) or on a national scale in the network of National Oceanographic Data Centres (NODCs) supported by the Intergovernmental Oceanographic Commission of UNESCO's International Oceanographic Data and Information Exchange

DOI: 10.4018/978-1-4666-8111-8.ch004

(IODE). This tradition led to the creation of several attempts to integrate the data from the DACs and NODCs either in portals or multinational data systems. Until recently, data were collected for one purpose, but research paradigms are shifting and expectations along with them for increased access to data and desire to integrate multidisciplinary data from disparate sources. While it is important to curate data close to the point of origin, it is important to 'set data free' and with sufficient documentation that the data can be interpreted accurately by others. A possible solution to providing sufficient documentation is through the use of Sensor Observation Services (SOS, http://www.opengeospatial.org/standards/sos). Fredericks (2015) presents a vision of semantic SOS, and this is to be extended in the ongoing European Commission SenseOCEAN project. We present here the results of a recent collaborative network in loosely coupling data systems using Semantic Web and Linked Open Data methodologies in order to generate new information and knowledge within the ocean science domain and in support of this new research paradigm, concentrating on the integration of multidisciplinary data from a range of sources. These methodologies are explained in detail by Leadbetter (2015).

Controlled vocabularies are collections of terms that may be used to populate a given element in a structured metadata document or a data file. The controlled vocabulary concept has been an essential part of the British Oceanographic Data Centre's (BODC) data infrastructure from its initial design in the late 1970s to the present day. However, the way in which controlled vocabularies have been managed has changed considerably, with development from loosely managed basic code tables into an integrated semantic resource with formal content governance and sophisticated serving technology based on internationally-agreed knowledge management standards. In this section of the chapter we will present the lessons that have been learned in building and maintaining these vocabularies through time and governing the content within communities.

In 2012, version 2.0 of the Natural Environment Research Council Vocabulary (NERC) Server (NVS) which is used to deliver these vocabularies was launched (Leadbetter, Lowry, & Clements, 2013b). This release underpins the rest of the work presented in this chapter.

Current practice in BODC controlled vocabulary management now underpins the semantic infrastructure of many oceanographic data management systems across the world, such as SeaDataNet in Europe (Schaap & Lowry, 2010), the Rolling Deck to Repository (R2R) programme (Arko, Chandler, Clark, Shepherd, & Moore, 2012) and the Biological and Chemical Oceanography Data Management Office (BCO-DMO) project (Chandler, Groman, Allison, Wiebe, & Glover, 2013) in the US. The NVS, BCO-DMO and R2R linkages are all implemented using Linked Open Data techniques (Leadbetter, Arko, Chandler, Shepherd, & Lowry, 2013a) which means that the collaboration has been driven through a "bottom-up" desire to integrate data sources to produce new knowledge, and not from a "top-down" imposition of standards. Recently, the Integrated Marine Observing System in Australia has begun a process of mapping its vocabularies to the NVS. A major benefit of this "bottom-up" approach is that we have maintained and respected the local naming conventions in use, and have created the integrated knowledge space through a mapping of naming terms, not a replacement. In this section we examine in detail the process of connecting the datasets and information sources: including the technical difficulties encountered and their solutions and the practices of both physical and virtual communication involved in creating an intercontinental network of this type. Other initiatives which have taken a similar approach include the Marine Metadata Interoperability project (http://marinemetadata.org/) which has its own Ontology Registry and Repository, with mappings from that resource to the NVS.