Reconstituting a Digital Repository through Use Case Driven Ontology Development

Conceptual modeling for the curation of oceanographic datasets

Adam Shepherd[†]
Marine Chemistry and Geochemistry
Woods Hole Oceanographic Institution
Woods Hole, MA USA
ashepherd@whoi.edu

Danie Kinkade[†]
Biology
Woods Hole Oceanographic Institution
Woods Hole, MA USA
dkinkade@whoi.edu

Douglas Fils[††]
Consortium of Ocean Leadership
Washington, D.C., USA
dfils@oceanleadership.org

ABSTRACT

The Biological and Chemical Oceanography and Data Management Office (BCO-DMO) was created in 2006 to assemble, curate, and publicly serve all data and related products resulting from grants funded by the NSF core programs for Biological and Chemical Oceanography, and Office of Polar Programs. Since that time, a rich database has been built of over 9,000 datasets of diverse data types from over 2,600 contributors from physical, chemical, biological and/or ecological and biogeochemical sub-domains representing 1,000 funded projects.

Due to changes in the data management landscape, BCO-DMO recently began to re-architect its data infrastructure with a focus on its conceptual model. The goal of this re-architecture was to simplify the management of its custom software to leverage the data management and knowledge representation expertise of its staff by migrating logical assumptions within code to an RDF ontology. With its schema and axioms, the ontology would serve as the backbone to a knowledge graph that would drive the functions and capabilities across BCO-DMO's software stack. Applying the Tetherless World Constellation (TWC) Semantic Web Methodology developed at the Tetherless World Constellation at Rensselaer Polytechnic Institute, BCO-DMO assembled a small team of various experts to redesign the BCO-DMO data infrastructure and in turn publish a new version of its Ocean Data Ontology.

Through iterating over concept maps, activity diagrams, and sequence graphs, BCO-DMO continues an iterative process of re-evaluating new and existing use cases for the next version of the Ocean Data Ontology. This paper will describe the process of centering a digital repository's infrastructure around its conceptual model detailing how the TWC Semantic Web Methodology was applied.

CCS CONCEPTS

• Digital libraries and archives • Entity relationship models
• Graph-based database models

KEYWORDS

Conceptual Models, Design Patterns, Ontologies, Data Repository, Resource Description Framework

1 Background

Scientific research is intrinsically reliant upon the creation, management, analysis, synthesis, and interpretation of data. Once generated, data are essential to demonstrating the veracity and reproducibility of scientific results, and existing data hold great potential to accelerate scientific discovery through reuse. The Biological and Chemical Oceanography and Data Management Office (BCO-DMO) is a publicly accessible earth science data repository created in 2006 to curate, serve (publish), and archive digital data and information from biological, chemical and biogeochemical research conducted in coastal, marine, great lakes and laboratory environments. The office's mission is to provide investigators with data management services that span the full data lifecycle (from data management planning, to data publication, and archive with an appropriate national facility). However, since 2006 data management has both significantly matured and grown progressively complex within the ocean sciences [1].

Best practices have emerged for data citation, versioning, provenance, and identification while big data challenges stress the capabilities of domain-specific repositories. These progressions required BCO-DMO re-architect its data infrastructure to address software development that had diverged from its conceptual model. This resulted in software that contained logical

[†]Adam Shepherd and Danie Kinkade are co-PIs of the Biological and Chemical Data Management Office (BCO-DMO) a National Science Foundation funded data repository. Together, they serve as the Technical Director and Director of the facility, respectively.
[††]Douglas Fils is the PI for the Open Core Data project who collaborated with BCO-DMO on its refactored conceptual model.

assumptions about the model that over time became expensive to manage. In effect, this valuable knowledge about how data and information were related was locked inside software unavailable to data producers, curators and consumers.

Despite working with ontologies since 2012, BCO-DMO used them solely to publish metadata following the W3C Best Practices for Data on the Web [2]. Yet, as knowledge graph technologies become increasingly useful to software development [3], BCO-DMO has recognized that ontology-driven software development can lower the technical debt incurred from having these logical assumptions and domain knowledge stored as software code. As a result, a major goal of the re-architecture was to leverage BCO-DMO's Ocean Data Ontology [4] to achieve a fully data-driven architecture using a common machine-accessible vocabulary across multiple software components. To redesign the data infrastructure and the Ocean Data Ontology, BCO-DMO applied the TWC Semantic Web Methodology [5].

2 Use Case Driven Approach

First introduced to the TWC Semantic Web Methodology (TWC-SWM) in 2010 by Dr. Peter Fox and the Tetherless World Constellation, a collaboration between Fox and BCO-DMO resulted in the first version of Ocean Data Ontology (ODO). After further training in the method at a 2014 workshop [6], the office has been successfully applying the technique [7] most recently on the Ocean Protein Portal project [8]. BCO-DMO decided to apply the method to its own data infrastructure to redesign ODO as a conceptual model for governing all functions of the repository from data submission to archive.

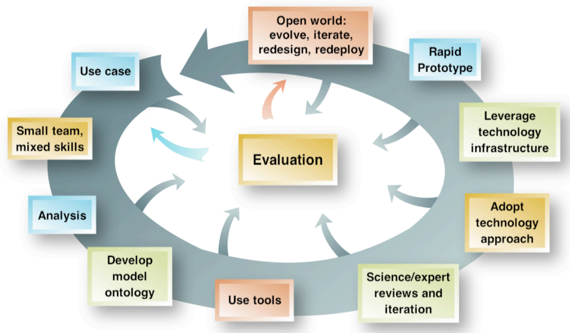

Open world: evolve, iterate, redesign, redeploy
Rapid Prototype
Use case
Leverage technology infrastructure
Small team, mixed skills
Evaluation
Adopt technology approach
Analysis
Develop model ontology
Use tools
Science/expert reviews and iteration

Figure 1. The iterative cycle of the TWC Semantic Web Methodology

TWC-SWM is an iterative process for developing semantic technologies that start with the construction of a use case within a small team of mixed experts (Figure 1). The team comprised of data managers with domain expertise in ocean science, knowledge engineers, informaticists and software engineers. To maximize efficiency, teams were grouped around the major functions of the repository - data submission, data processing, metadata curation, knowledge acquisition, and data serving/access.

For each major function, teams developed use cases to capture the desired result. In the TWC-SWM, the use case should identify the question(s) that will be addressed by that use case and the resources required to answer those questions [9]. For example, one of our questions for the data processing function was, "How do we capture provenance for each action performed by a BCO-DMO data manager on a submitted dataset?" The process of answering this question begins by identifying the actors and resources involved, what triggers the use case, and any expected pre- and post-conditions. Typically, use cases are captured as documents using templates [10] that contextualize the use case's questions. Apart from identifying the actors in the use case, triggers, and conditions, these documents also contain prompts for activity diagrams, and step-by-step descriptions of the logical flows during the execution of the use case. With a conglomerate of use cases across major functions of the repository, BCO-DMO was able to aggregate knowledge across those use cases to develop a new conceptual model for the Ocean Data Ontology.

3 A New Ocean Data Ontology

When iterating through TWC-SWM, developing a shared vocabulary within the team working on a use case becomes critical. Terms within that vocabulary should be identified and defined to ensure that all team members agree what is being referenced when a term is used. Since ontology development provides a mechanism to capture these terms, it would be logical to start creating an ontology at this stage. However, the TWC-SWM suggests that working at the conceptual level during information modeling is most effective [5]. After reviewing the use cases developed across the major functions of the repository, the team working on the Ocean Data Ontology used simple concept maps to quickly grasp the relationships between resources in the data model (Figure 2). The driving question of the ODO redesign use case became, '*What is the knowledge model required for capturing the data and information necessary to execute the use cases of BCO-DMO's major functions?*'

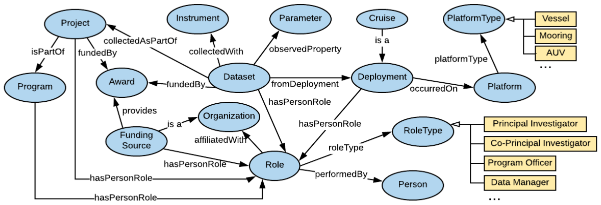


Figure 2. An early, simplified concept map for communicating the metadata resources BCO-DMO curates to those unfamiliar with the repository.

The simplified concept map was broken down into more detailed, smaller concept maps that dove deeper into a single concept (Figure 3). While the overall concept map identified the major concepts involved, the deeper concept maps fleshed out all properties of a single concept and followed a convention for graphical notation that was easily translated into an ontology (Figure 4).

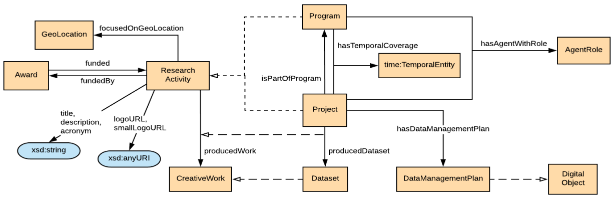


Figure 3. A concept map of a Research Activity and its subclasses – Project and Program.

As the concept maps are evaluated and iterated, they are encoded into the Ocean Data Ontology using the Protege tool. Through Web Ontology Language (OWL) annotation properties, these concept maps are explicitly referenced from the corresponding class in the ontology (Figure 5). Figure 5 depicts a snapshot of the annotation property, odo:lucidchartDiagram, defining the URL to where the concept map for the ResearchActivity class can be found. Finally, the ontology is published to the web at the ontology's base URI http://ocean-data.org/schema/ using WIDOCO [11] to supply RDF and HTML versions of the ontology in following the W3C Best Practices for Linking Data [12].

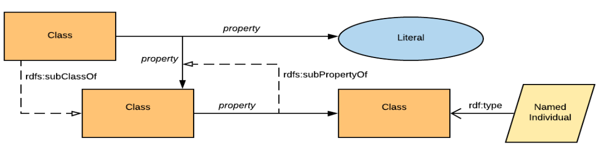
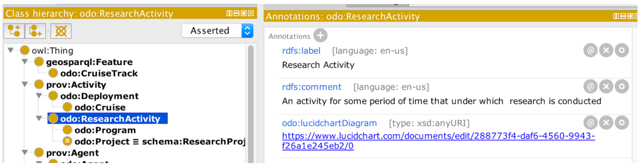


Figure 4. The graphical notation used at BCO-DMO for describing classes, properties, literal values and named individuals.


Class hierarchy: odo:ResearchActivity
Asserted
Annotations: odo:ResearchActivity
Annotations

owl:Thing
geosparql:Feature
odo:CruiseTrack
prov:Activity
odo:Deployment
odo:Cruise
odo:ResearchActivity
odo:Program
odo:Project ≡ schema:ResearchProj
prov:Agent

rdfs:label [language: en-us]
Research Activity

rdfs:comment [language: en-us]
An activity for some period of time that under which research is conducted

odo:lucidchartDiagram [type: xsd:anyURI]
https://www.lucidchart.com/documents/edit/288773f4-daf6-4560-9943-f26a1e245eb2/0

Figure 5. Annotating the Research Activity class in the Ocean Data Ontology with the URL to its corresponding concept map.

4 Conclusion

Needing to re-architect its digital repository to meet new challenges of big data and distributed resources, BCO-DMO sought to forge a sustainable repository by governing its software and information through conceptual model development. Through the TWC Semantic Web Methodology, BCO-DMO developed use cases that help transform mental maps across a team of experts into actionable models that will result in a new version of the Ocean Data Ontology in the coming months. This technique can be iteratively applied to avoid previous problems of having logical assumptions embedded in software. BCO-DMO has an operational development cycle that evaluates new needs of the data infrastructure, maps and models those needs into a publicly-available ontology that can be used to describe data and information as it moves through the data management life cycle.

ACKNOWLEDGMENTS

BCO-DMO is funded by NSF Award #1435578. We would like to thank Peter Fox, co-chair of the Tetherless World Constellation at the Rensselaer Polytechnic Institute, for introducing BCO-DMO to the Semantic Web Methodology.

REFERENCES

[1] Kinkade, D; CL Chandler; RC Groman; A Shepherd; MD Allison; S Rauch; PH Wiebe; DM Glover (2014) "Navigating a Sea of Big Data" (poster) Abstract IN11C-3622, presented at 2014 Fall Meeting, AGU, San Francisco, CA, 15 December 2014.

[2] Lóscio, B. F., Burle, C., & Calegari, N. (2017). Data on the Web Best Practices, W3C Recommendation 31 January 2017. Retrieved April 23, 2019, from https://www.w3.org/TR/dwbp/

[3] Uschold, M., Ding, Y., & Groth, P. (2018). Demystifying Owl for the Enterprise. Morgan & Claypool Publishers. ISBN:1681731274 9781681731278

[4] Shepherd, A. (2018). BCODMO/Ocean-Data-Ontology. Zenodo. doi:10.5281/zenodo.1285187

[5] Fox, P. and McGuinness, D.L. (2008). TWC Semantic Web Methodology. Retrieved April 15, 2019, from https://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology.

[6] Lightstrom, F. (2014). Use-Case Training for the Woods Hole, Massachusetts, Research Community. Retrieved April 15, 2019, from https://soundwaves.usgs.gov/2014/08/meetings2.html.

[7] Leadbetter, A., A. Shepherd, R. Arko. (2016) Experiences of a "semantics smackdown". Earth Science Informatics; volume 9, issue 3, pg 355-363. doi:10.1007/s12145-016-0252-8

[8] Saito, M., J. Saunders, N. Held and A. Shepherd (2019). Introducing an Ocean Protein Portal Town Hall. Presented at the ASLO 2019 Ocean Sciences Meeting. San Juan, Puerto Rico. February 23 - March 2, 2019.

[9] Ma, X., J. Guang Zheng, J. Goldstein, S. Aulenbach, C. Tilmes, Curt and P. Fox. (2013). A use case-driven iterative method for building a provenance-aware GCIS ontology. Presented at the 2013 ESIP Summer Meeting. Chapel Hill, NC. July 9-12, 2013.

[10] Cockburn, A. (2000). Writing Effective Use Cases (1st ed.). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

[11] Garijo, D., M. Scharm, A. Ruiz-Iniesta, J. Geluk, kartgk, María, O. Corcho, M. Angel Garcia, M. Lefrançois and J. Schneider. (2019). dgarijo/Widoco: WIDOCO 1.4.9: Supporting schema.org (Version v1.4.9). Zenodo. doi:10.5281/zenodo.2576182

[12] Hyland, B., G. Atemezing and B. Villazón-Terrazas (2014). Best Practices for Publishing Linked Data. W3C Working Group Note 09 January 2014. Retrieved April 23, 2019, from https://www.w3.org/TR/ld-bp/