The OceanLink Project

Tom Narock
Department of Information Technology
and Management Science
Marymount University
Arlington, VA, USA

Robert Arko, Suzanne Carbotte
Lamont-Doherty Earth Observatory
Columbia University
New York, NY, USA

Adila Krisnadhi, Pascal Hitzler, Michelle Cheatham
Department of Computer Science
Wright State University
Dayton, OH, USA

Adam Shepherd, Cynthia Chandler,
Lisa Raymond, Peter Wiebe
Woods Hole Oceanographic Institution
Woods Hole, MA, USA

Timothy Finin
Department of Computer Science
University of Maryland, Baltimore County
Baltimore, MD, USA

***Abstract*—Today's scientific investigations are producing large numbers of scholarly products. These products continue to increase in diversity and complexity as researchers recognize that scholarly achievements are not only published articles but also datasets, software, and associated supporting materials. OceanLink is an online platform that addresses scholarly discovery and collaboration in the ocean sciences. The OceanLink project leverages Semantic Web technologies, web mining, and crowdsourcing to identify links between data centers, digital repositories, and professional societies to enhance discovery, enable collaboration, and begin to assess research contribution.**

Keywords—Semantic Web; ocean science; Big Data; Ontology Design Patterns

I. INTRODUCTION

Today's scientific investigations are producing large numbers of scholarly products. These products continue to increase in diversity and complexity as researchers recognize that scholarly achievements are not only published articles but also datasets, software, and associated supporting materials. Yet, discovery of these related resources is still a daunting task. Many researchers are unfamiliar with all the resources that are available to them. Relationships amongst these resources are implicit and difficult, if not impossible, to search computationally.

OceanLink is an online platform that addresses scholarly discovery and collaboration in the ocean sciences. A wide spectrum of maturing methods and tools, collectively characterized as the Semantic Web, have the ability to vastly improve the discovery and dissemination of scientific research. OceanLink leverages the Semantic Web, in conjunction with web mining and crowdsourcing, to identify links between data centers, digital repositories, and professional societies. This is leading to new search capabilities for scholarly discovery and collaboration. OceanLink is also producing new analytical tools to assess the impact of research funding and scholarly work. Moreover, OceanLink must scale to operate in the era of Big Data. The current OceanLink platform contains tens of millions of semantic statements. The final platform is anticipated to approach one billion statements.

II. BACKGROUND

In 2011 the National Science Foundation (NSF) launched EarthCube, an initiative with the goal of creating a more sustainable future through improvements in understanding Earth and our changing planet. As a cornerstone of NSF's Cyberinfrastructure for the 21st Century (CIF21), EarthCube's goal is to be reached through the development of community-based, sustainable, and nationwide cyberinfrastructure. What this cyberinfrastructure should look like was assessed through several domain workshops.

The *Needs of the Ocean Ecosystem Dynamics Community Workshop* [1] was of particular interest to our ocean sciences application domain. Held at the Woods Hole Oceanographic Institution in October of 2013, this workshop had the goal of articulating cyberinfrastructure needs of the ocean ecosystem dynamics community with particular focus on the challenges presented by multi-disciplinary marine ecosystem research. The ocean ecosystem dynamics domain encompasses a broad array of disciplines that seeks to increase understanding of the interplay between biological, chemical, and physical processes in the ocean. "It is fundamentally an interdisciplinary domain by nature, producing highly diverse data types that pose unique challenges for management, integration, and analysis. The ability to discover, access, and synthesize high quality data from various disciplines is crucial to ocean ecosystem sciences" [1].

978-1-4799-5666-1/14/$31.00 ©2014 IEEE

What emerged from this workshop, as well as other NSF supported domain workshops, is that a lack of interoperability exists among data centers as well as a lack of rich, standardized metadata. To address these challenges an initial set of EarthCube Building Blocks was funded in 2013. OceanLink was among those first Building Blocks with the goal of applying semantic technologies to the data discovery and integration problem, with the ocean sciences as an exemplar.

OceanLink is demonstrating semantic technologies through the integration of ocean science data repositories, library holdings, conference abstracts, and funded research awards. Through OceanLink we are applying semantic technologies to support data representation, discovery, sharing and integration. We have developed and tested prototype semantic cyberinfrastructure components in support of the geoscience research community. Here we report on both the technical and social implications of this work. We discuss recent advances in semantic technologies and methodologies and look at how this impacts OceanLink's architecture and usage. We present newly published open data and ontologies in the ocean sciences and give examples of OceanLink's usage. In doing so, we highlight how semantic technologies enable knowledge discovery and integration. We also demonstrate how these interlinked resources can be leveraged to assess scholarly impact as well as enhance collaboration.

III. Related work

Geoscience researchers are actively working toward a research environment of software tools and interfaces to data archives with the goal of full-scale semantic integration beginning to take shape. The geoscience community, early adopters of semantic technologies, have provided essential feedback to the semantic web community [2]. Yet, growth in semantic geoscience and the more general computer science Semantic Web community have continued largely independent of each other [2, 3]. Moreover, achieving scalable and sustainable semantic integration requires both domain and cyberinfrastructure scientists and contains both social and technical aspects.

Neumann [4] began looking at the Semantic Web in the life sciences nearly a decade ago and Fox et al. [5] soon followed with a system that utilized semantic technologies in the discovery and integration of solar-terrestrial data. Subsequent works have applied semantic technologies in ecology [6] and Heliophysics [7]. And additional studies [8] have shown the benefits of semantic systems over non-semantic systems in data discovery. Today, a significant portion of the "Web of Data" is now comprised of science data and scientists stand to benefit from the afforded data fusion capabilities [9]. Specifically in the area of ocean sciences, there exists semantic encodings of marine data [10], oceanographic expeditions [11], and biological and chemical oceanography data [11].

Yet, the independent growth of the geoscience and computer science communities [2, 3] means that emerging semantic web technologies and methodologies have had limited usage in the geosciences in which to evaluate their large scale applicability. In addition, the applications that do exist focus exclusively on scientific data discovery and integration. OceanLink takes the next step and begins to link data to other scholarly resources.

A. Ontology Design Patterns

One such emerging methodology is that of Ontology Design Patterns (ODPs) [12]. ODPs enable horizontal integration between repositories with potentially independent semantic models. This is opposed to the traditional approach advocating an overarching upper-level ontology that captures global agreement on concepts, something that is often infeasible even within a single scientific domain [13]. Instead, the ODP approach advocates a set of partial ontologies, each of which formalizes only one key notion. The axiomatization of each pattern (partial ontology) encodes what constitutes the given notion and what domain experts have agreed upon, with a focus on eliminating application specific semantic statements. In this manner, data providers can horizontally align key notions by mapping their content to supplied ODPs. A data provider can align to one or more ODPs as needed. This is in contrast to upper-level ontologies in which a user must accept all of the semantics and entailments that come from the monolithic ontology. ODPs can thus provide partial integration in ways that upper-level ontologies cannot.

Working with computer scientists and domain experts we have created a set of ODPs relevant to the ocean science community. The patterns are publicly available [1] and will continue to grow as the OceanLink project evolves. These patterns provide the basis of our semantic integration and are used to provide federated queries amongst our constituent data providers. In addition to the ODPs themselves, the OceanLink project has produced an open source software library [2] that produces semantic metadata compliant with the ODPs.

IV. OceanLink Architecture

A. Data Sets and Community Contributions

The primary goal of OceanLink is the use of semantic technologies to support data representation, discovery, sharing and integration. To this end, we have identified leading data repositories and related contextual and supporting data sets for our initial integration. This includes ocean science data repositories, library holdings, conference abstracts, and funded research awards. Specifically, we are working with the Biological and Chemical Ocean Data Management Office [3] (BCO-DMO), which serves data from research projects funded by the Biological and Chemical Oceanography Sections and the Division of Polar Antartic Organism & Ecosystems Program at the U. S. National Science Foundation. OceanLink also contains access to research vessel and cruise descriptions, inventories of original environmental sensor data sets, quality assessment reports, and standard trackline navigation products from the Rolling Deck to

[1] http://schema.oceanlink.org/

[2] https://github.com/narock/earthcube-EAGER

[3] http://www.bco-dmo.org/

Repository [4] (R2R) initiative. These ocean data are complimented with links to relevant resources at the Marine Biological Laboratory Woods Hole Oceanographic Institution Library[5] (MBLWHOI), conference abstract and session data from the American Geophysical Union [6] (AGU) Ocean Science and Fall meetings as well as funded awards from the National Science Foundation's historical funded project database.

The aforementioned ODPs were deployed at each data provider's site to enable horizontal integration. There are two important points to note regarding the application of ODPs. First, only relevant ODPs need to be applied allowing for ease of implementation and partial alignment. For example, while R2R has *Event*s (research cruises), *Vessel*s, and *Cruise*s, the AGU conference abstracts contain only one overlapping term – *Event*. The use of ODPs allows us to reuse the semantics of *Event*, providing alignment of this notion, without importing additional semantics and entailments as is often the case with overarching upper-level ontologies. Second, data providers are only required to produce semantic metadata and do not have to change existing data production and distribution methods. In some cases existing metadata can be automatically mapped to ODP semantic metadata. Open source projects such as D2RQ[7] enable the mapping of relational databases to Semantic Web standards. This is the process used by R2R. The OceanLink portal, described in section D, distributes user queries to the relevant participating data repositories. The results are aggregated by the portal and a unified picture of the ocean science domain is presented to the user.

Achieving scalable and sustainable semantic integration requires domain and cyberinfrastructure scientists, a process that is both technical and social. Many domain data providers are unfamiliar with semantic technologies and their proper usage and setup can be a daunting task. Yet, the power of semantic technologies is human and machine understandable data and we can leverage this to lower the barrier to entry.

The World Wide Web Consortium's (W3C) Vocabulary of Interlinked Datasets[8] (VoID) is a vocabulary for specifying metadata about semantic datasets. VoID provides the bridge between publishers and users. One of the many uses of VoID is to describe the semantic concepts that have been used in a published data set. OceanLink creates VoID semantics to formally specify which semantic concepts are available at each of the constituent data repositories. The creation of semantic metadata conforming to the ODPs and the creation of a VoID description constitutes registration of a data provider into the system. Using the VoID descriptions the OceanLink portal is able to determine which concepts a data provider is claiming it knows about. With the Semantic Web query language SPARQL the portal can query each data provider and test if it responds properly to each concept it claims to know about. For example, R2R claims to support *Cruise* queries and OceanLink can validate this claim through several pre-defined *Cruise* queries. The OceanLink portal contains a page devoted to automatically reading the VoID descriptions, querying data repositories, and presenting the results in an easy to read table. In this manner, data providers can easily and quickly assess the validity of their implementations. Figure 1 shows a graphical example of this from the early implementation of OceanLink.

Current status of horizontal alignment using ontology design patterns.

N/A indicates that a repository does not currently support the concept(s). Green "Ok" indicates that the repository supports the concept(s) and is responding properly to SPARQL queries. Red "No Results" indicates the repository claims to support the concept, but is currently not responding properly to SPARQL queries. Clicking on the Concept name will show the SPARQL query that is being submitted

Concept	AGU	BCO-DMO	R2R	WHOI-Library
Cruise and Vessel	N/A	Ok	Ok	N/A
Person (First and Last Name)	No Results	Ok	No Results	Ok
Person (Full Name as String)	No Results	Ok	No Results	No Results
Program	N/A	Ok	Ok	N/A
Repository Objects	Ok	N/A	N/A	No Results

Figure 1. Using semantic technologies the OceanLink portal assess the current state of alignment amongst participating facilities. This image shows the results of this automated quality assessment during the development of OceanLink.

The VoID metadata can additionally be utilized to optimize OceanLink. For a given facet, OceanLink can utilize the VoID semantics to determine which repositories should be queried. This avoids unnecessarily broadcasting all queries out to all data providers. Such a system would not be scalable. Figure 2 highlights a simple evaluation of this. Given an increasing number of data providers we seek to simulate the delay caused by unnecessarily querying non-relevant data providers. In this simulation we assume there are two providers that have metadata relevant to a query. We look at how long it would take to blindly query all data providers as opposed to looking at VoID semantics and only querying relevant providers. Figure 2 highlights the results of this simulation. The blue line (diamond points) shows response times when the query is distributed to all providers while the red line (square points) shows the response times when only the two relevant providers are queried. Evident from this figure is that a system that blindly queries all providers quickly introduces unnecessary delays to the user. This is particularly important given OceanLink's intentions to expand to numerous other geoscience repositories and roughly one billion semantic statements.

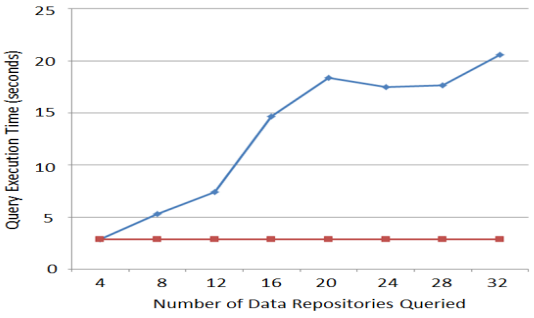


Figure 2. Query execution time as a function of number of data repositories queried.

[4] http://www.rvdata.us/overview/
[5] http://mblwhoilibrary.org/
[6] http://abstractsearch.agu.org/about/
[7] http://d2rq.org/
[8] http://www.w3.org/TR/void/

The OceanLink project has also contributed to the open source software community in an attempt to ease entry into semantic technologies. Drupal[9] is an open source content management system commonly used within the Earth sciences. Members of the OceanLink team have extended two popular Drupal modules[10,11] to work with the RDF database Virtuoso. These extensions have been contributed back to the Drupal community such that additional data providers will have easier access to OceanLink in particular and Semantic Web Technologies in general.

DSpace[12] is analogous to Drupal in that it is an open source repository package. However, DSpace is typically used for creating open access repositories for scholarly content. It differs from Drupal in that DSpace has specific functionality focused on digital archives and the storage, access, and preservation of digital content. Because of this, DSpace is popular amongst many institutional repositories. Working with our library colleagues we were able to create new DSpace extensions that create semantic metadata. Some of this semantic metadata is specific to the OceanLink project; yet, this extension also produces generic Dublin core semantic metadata applicable to many semantic projects.

B. Data Acquisition and Provenance

The OceanLink portal provides an aggregated view of the ocean sciences domain. Yet, it is not a clearinghouse containing all ocean science resources. The OceanLink platform, of which the portal is the user facing web interface, contains a set of algorithms for the automated identification of links between resources. This includes such things as continual web mining of AGU conference abstracts for mentions of cruises and vessels as well as algorithms for identifying the same entities mentioned differently across data sets. The latter is commonly referred to entity resolution and occurs, for instance, when the AGU has a record for *P. Wiebe* and ocean science data repositories refer to *Peter Wiebe*. Reconciling that these are indeed the same person is a challenging and outstanding problem in computer science.

In the context of OceanLink, entity resolution enables the accurate validation of an entity's property values - an exercise highly valued by data management practitioners. BCO-DMO semantically models metadata surrounding oceanographic research cruises, but other sources outside of BCO-DMO exist that also model metadata about these same cruises. The R2R program presents a data context that is uniquely different and valuable. Where BCO-DMO exposes the processed and analyzed data products generated by researchers, R2R exposes the raw shipboard data that led to those processed and analyzed data products. Identifying that these products came from the same cruise expands data discovery capabilities, but also allows for validating the contextual correctness of both BCO-DMO and R2R cruise metadata. Assessing the potential for a link between two cruises consists of aligning like properties and deciding on the appropriate semantic markup to describe the link. This highlights the desire for research organizations like BCO-DMO and R2R to ensure the complete accuracy of research metadata – the more accurate the metadata the more easily other repositories can identify links.

Thus, the OceanLink entity resolution algorithms are being built to not only identify similar entities, but also to also assist data managers with quality control and quality assurance. For instance, one algorithm [14] uses text similarity (Levenshtein distance) to compare the property values of ontology instances. For a research cruise this would include comparing values such as cruise start date, cruise end date, vessel name, and chief scientist. These property similarity values are then weighted and aggregated to compute an overall match confidence score between the two entities. Appropriate property weights and threshold match confidence score (>= 85% indicates a match) have been determined through simulation experiments. All output of this algorithm (entities being compared, overall match confidence, and individual property similarity values) is stored as semantic provenance information using the W3C PROV-O recommendation [15]. Software tools can then be built to leverage this provenance information. One such tool, shown in Figure 3, uses the entity resolution provenance to show data managers which repositories contain what the algorithm assumes is the same entity, but with different property values. This side-by-side comparison helps data managers quickly assess potential errors in their metadata.

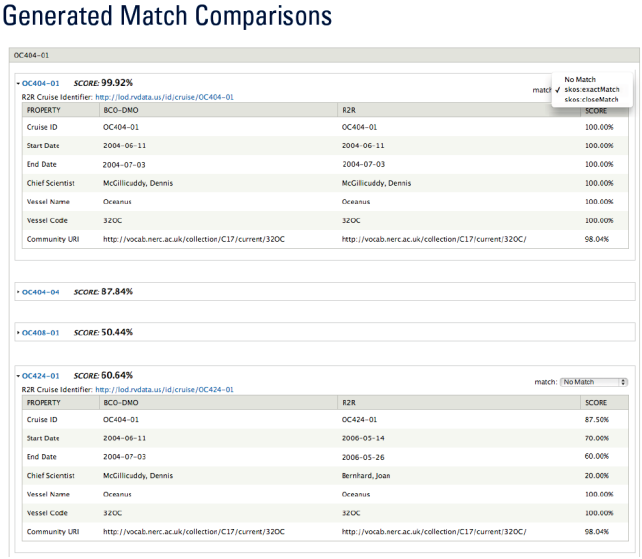


Figure 3. Sample screen shot of quality control tool built on top of entity resolution and provenance data. Reproduced with permission from [14].

A second entity resolution tool being explored [16] leverages ensemble learning techniques. In this approach, entity resolution is determined via a collection of approaches that currently includes Point Wise Mutual Information, Support Vector Machine, and K-Nearest Neighbor. If the ensemble classifiers agree that two entities are identical then it

[9] https://www.drupal.org/
[10] https://www.drupal.org/project/rdfx
[11] https://www.drupal.org/project/rdf_indexer
[12] http://www.dspace.org

is recorded as a match at the OceanLink portal. If the classifiers do not agree, the entity pair is sent out for crowdsourcing to domain experts. Each domain expert is offered a thumbs up (entities are the same) thumbs down (entities are distinct) vote with the majority determining the outcome. The ensemble approach helps ensure that human computation is only leveraged when absolutely necessary.

The OceanLink portal must, at least temporarily, have access to the collective metadata in order for these algorithms to run. Similar efforts in the Semantic Web community (e.g. [17, 18]) simply download the entire holdings of participants at regular intervals. Yet, using the Semantic Web query language SPARQL to periodically download all data in the era of Big Data will not scale. Instead, when applicable, OceanLink has been exploring a process based on recent advances in semantic provenance [15]. By extending semantic metadata with provenance the portal is able to download only what is new or modified. We are experimenting with the PROV-O [15] notion of a *Collection* and the DESCRIBE feature of the SPARQL query language to significantly reduce the amount of provenance information that needs to be created and stored. As a simple example, the use of provenance allowed the OceanLink algorithms to infer that only 21,296 of the 263,434 AGU conference abstracts were new (from the most recent 2013 meeting) and needed to be mined.

C. Crowdsourcing

In addition to the aforementioned ensemble learning approach, we are also aware that our text mining system will inevitably identify incorrect links. Moreover, there are some types of links that can currently be identified only by human computation. For example, the geosciences currently lack a consistent method for citing a dataset [19]. Approaches vary across federal research centers such as NASA and NOAA, and in some cases vary between types of data. Even when recommendations are available, nuances such as dataset version number and processing level make automated identification intractable. Thus, identification of links between publication and datasets used is currently a manual task within the geosciences.

To this end, we have begun exploring a comprehensive crowdsourcing [20] platform. All of the identified links, both through computational reasoning and text mining, are stored on the OceanLink portal. We are integrating a crowdsourcing platform into OceanLink that will allow the community to validate these links as well as create new ones. In doing so, we are creating a feedback loop in which machine computation is validated, and extended, by human computation. This validation information, as well as user attribution provenance, will be folded back into the system to continually increase the precision of OceanLink.

D. OceanLink Portal and Provenance

Semantically aligned resources are of little use if users do not know where to find them. Indeed, one of the significant findings from NSF domain workshops, and the ocean sciences in particular [1], is that end-users are not aware of all the domain resources available to them. To alleviate this we created the OceanLink portal[13] – a one-stop place for querying our providers.

Using various facets users can create queries that are then distributed amongst the constituent data providers. The end-user need not know which resources exist or where they are located. The portal queries all relevant providers and aggregates the results for the user. An example scenario is shown in Figures 4 and 5.

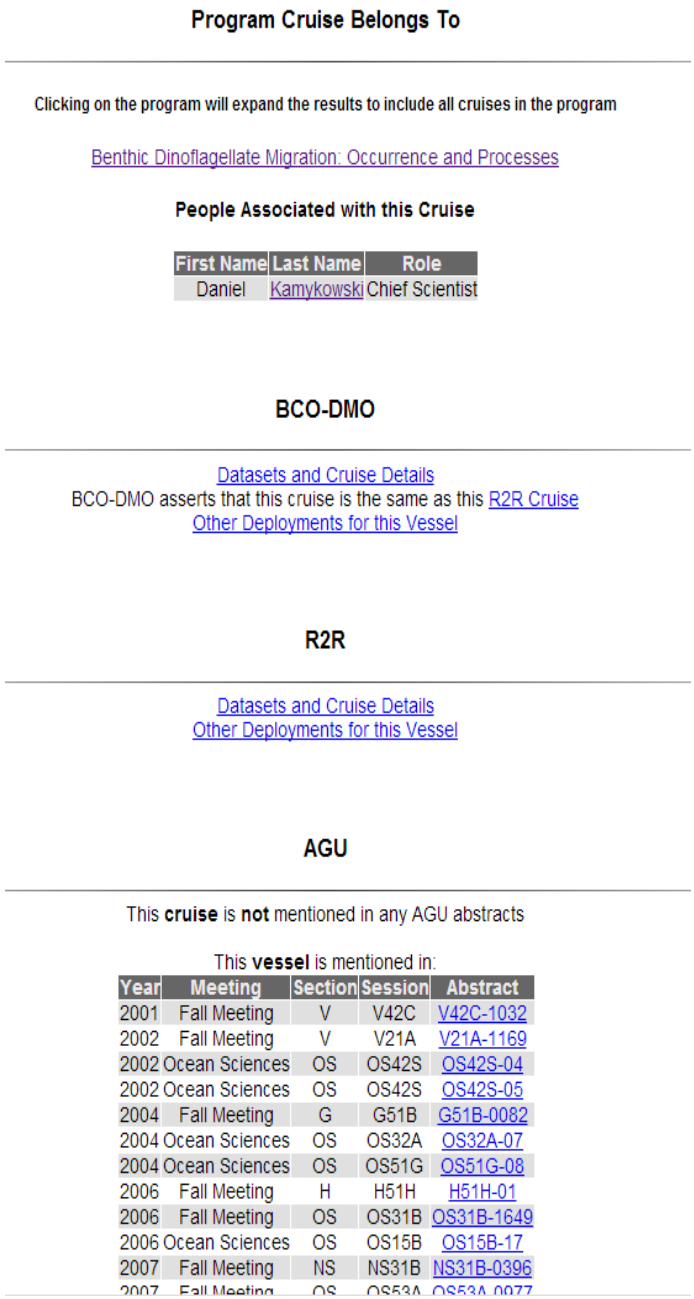
Cruise ID	Vessel Name
PEJun2000	R/V Pelican
PS02_2002	USCGC Polar Star
PS0819	R/V Point Sur
PS1009	R/V Point Sur
PacFlux_II_cruise_2	R/V John V. Vickers
Palau_reefs_2011-12	PICRC Small Boats
Pfister_2009	Pfister small boat
Pi_64_442	R/V Pioneer

Figure 4. Snippet of the OceanLink portal interface showing some of the results for querying all Cruise IDs beginning with "P".

Currently available search facets are Cruise ID, Program, and Person. Figure 4 shows a snippet of the OceanLink portal interface. In this example the user has selected Cruise ID and asked for all cruises beginning with the letter "P". Of note is that nowhere in this process did the user need to specify data repositories to query. Rather, the portal inferred all relevant data providers (those that support the *Cruise* concept), submitted the query, and aggregated the results to produce the table in Figure 4.

Continuing our example, let's suppose the user then selects cruise PE10-01. For brevity, the complete list of "P" cruises was not shown in Figure 4. Figure 5 shows the results page for cruise PE10-01.

[13] http://www.oceanlink.org



Program Cruise Belongs To

Clicking on the program will expand the results to include all cruises in the program

Benthic Dinoflagellate Migration: Occurrence and Processes

People Associated with this Cruise

First Name	Last Name	Role
Daniel	Kamykowski	Chief Scientist

BCO-DMO

Datasets and Cruise Details
BCO-DMO asserts that this cruise is the same as this R2R Cruise
Other Deployments for this Vessel

R2R

Datasets and Cruise Details
Other Deployments for this Vessel

AGU

This **cruise** is **not** mentioned in any AGU abstracts

This **vessel** is mentioned in:

Year	Meeting	Section	Session	Abstract
2001	Fall Meeting	V	V42C	V42C-1032
2002	Fall Meeting	V	V21A	V21A-1169
2002	Ocean Sciences	OS	OS42S	OS42S-04
2002	Ocean Sciences	OS	OS42S	OS42S-05
2004	Fall Meeting	G	G51B	G51B-0082
2004	Ocean Sciences	OS	OS32A	OS32A-07
2004	Ocean Sciences	OS	OS51G	OS51G-08
2006	Fall Meeting	H	H51H	H51H-01
2006	Fall Meeting	OS	OS31B	OS31B-1649
2006	Ocean Sciences	OS	OS15B	OS15B-17
2007	Fall Meeting	NS	NS31B	NS31B-0396
2007	Fall Meeting	OS	OS53A	OS53A-0977

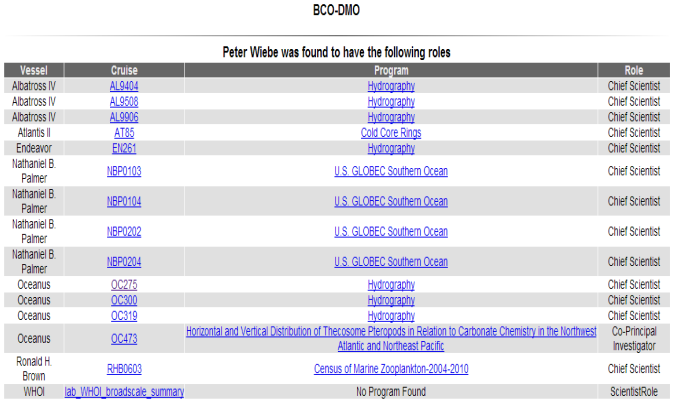
Figure 5. OceanLink search results for Cruise ID PE10-01

Evident from Figure 5 is that the OceanLink portal aggregated the results from several providers on the user's behalf. In addition, the semantics were leveraged to identify other relevant links. For example, we note that BCO-DMO and R2R both had metadata on this cruise. Further, the system asserts that the BCO-DMO cruise metadata and R2R's metadata are referring to the same cruise. Something that is not always easy for users to do manually given variations in local data repository identifiers. OceanLink's computational reasoning was able to infer that Cruises take place on Vessels and use this information to provide links to other deployments of the PE10-01's vessel. The portal also followed additional semantic links at the providers to identify the Program to which this cruise belonged and infer the other cruises that are also part of that same program. Finally, through text mining, the vessel name was identified to have appeared in the text of numerous AGU conference abstracts. All of this information is conveniently and uniformly aggregated and presented to the user.

V. Discussion

Semantic technologies leverage a graph-based data storage format. Our OceanLink graph can be exploited from various perspectives to enable collaboration as well as discover the impacts of scholarly works. In the preceding section, we showed an example of identifying additional scholarly products relevant to an oceanographic cruise. Those scholarly products included the traditional datasets and conference proceedings. The OceanLink team is actively working with the MBLWHOI Library to link to additional library resources such as student theses, cruise notes, and ship logs. In a similar manner, we are integrating the NSF historical funded project database into the system. This will enable linkages, for example, between the grant that funded a research cruise and all of the subsequent datasets, publications, and other products that resulted from that cruise. This will enable the regular tracking of scholarly impact. A preliminary example of this can be accomplished with the current OceanLink implementation.

If we were to choose the "Person" facet instead of the "Cruise ID" facet, we could, for example, search for the oceanographer Peter Wiebe. Results of this search are shown in Figure 6.



BCO-DMO

Peter Wiebe was found to have the following roles

Vessel	Cruise	Program	Role
Albatross IV	AL9404	Hydrography	Chief Scientist
Albatross IV	AL9508	Hydrography	Chief Scientist
Albatross IV	AL9906	Hydrography	Chief Scientist
Atlantis II	AT85	Cold Core Rings	Chief Scientist
Endeavor	EN261	Hydrography	Chief Scientist
Nathaniel B. Palmer	NBP0103	U.S. GLOBEC Southern Ocean	Chief Scientist
Nathaniel B. Palmer	NBP0104	U.S. GLOBEC Southern Ocean	Chief Scientist
Nathaniel B. Palmer	NBP0202	U.S. GLOBEC Southern Ocean	Chief Scientist
Nathaniel B. Palmer	NBP0204	U.S. GLOBEC Southern Ocean	Chief Scientist
Oceanus	OC275	Hydrography	Chief Scientist
Oceanus	OC300	Hydrography	Chief Scientist
Oceanus	OC319	Hydrography	Chief Scientist
Oceanus	OC473	Horizontal and Vertical Distribution of Thecosome Pteropods in Relation to Carbonate Chemistry in the Northwest Atlantic and Northeast Pacific	Co-Principal Investigator
Ronald H. Brown	RHB0603	Census of Marine Zooplankton-2004-2010	Chief Scientist
WHOI	lab_WHOI_broadscale_summary	No Program Found	ScientistRole

Figure 6. A search for the oceanographer Peter Wiebe reveals his research cruises and his role on those cruises.

From these results we can see that Peter Wiebe had the role Chief Scientist on a number of oceanographic research cruises. The Chief Scientist oversees the operations of a cruise and is responsible for the production of data sets from that cruise. Choosing the cruise NBP0103 we are led to the results shown in Figure 7 showing that this particular cruise was mentioned in two Ocean Science Conference abstracts. We are also given links to those conference abstracts.

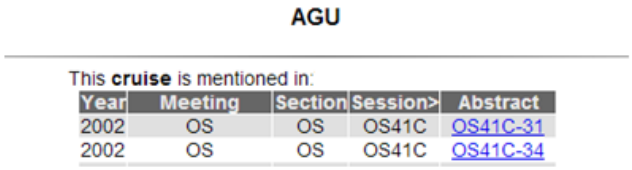


Figure 7. A snippet of the search results for cruise NBP0103 showing that this cruise was mentioned in two conference abstracts.

This simple example highlights the potential of OceanLink in particular and semantic technologies in general. Through a combination of machine reasoning and web mining we are beginning to link disparate scholarly resources within a single web platform. Given our initial datasets we can already start to see the scholarly endeavors, such as leading a cruise, are having by following links to additional resources. The depth and breadth of these links give us insights into the reach of one's work. These capabilities will continue to grow as additional digital repositories and scholarly products are added to the system.

VI. Conclusions

Semantics have been utilized in the geosciences for roughly a decade. During this time the geoscience community has been willing early adopters of semantic technologies and have provided essential feedback to the semantic web community [2]. Yet, growth in semantic geoscience and the more general computer science semantic web community have continued largely independent of each other [2, 3].

This work attempts to bridge that gap by once again using the geoscience as an early adopter of semantic technologies. The emerging semantic methodology of Ontology Design Patterns shows promise in large-scale semantic integration and OceanLink is among the first large-scale geoscience applications of this methodology. The OceanLink project is validating the ODP paradigm and showing increased scalability over traditional sematic integration, namely upper ontologies. In doing so, OceanLink is providing important feedback to the semantic web community.

The OceanLink project has also investigated the socio-technical aspects of semantic integration. We have released open source software libraries to aid in the creation of semantic metadata as well as automated the validation of a data provider's implementation. OceanLink has also broadened semantic integration in the geosciences. We are not only interlinking datasets, but also additional scholarly resources such as publications, library holdings, and funded awards. In the end, this is all enabling easier discovery of geoscience data and resources. OceanLink is currently scaling out to tens of data providers and an estimated one billion semantic statements. In doing so we are providing a building block for future semantic geoscience integration that enhances discovery, enables collaboration, and begins to assess research contributions.

References

[1] D. Kinkade, C. Chandler, D. Glover, R. Groman, D. Kline, J. Nahorniak, T. O'Brien, M. J. Perry, J. Pierson, and P. Wiebe, Articulating Cyberinfrastructure Needs of the Ocean Ecosystem Dynamics Community, Executive Summary of EarthCube End-User Domain Workshop held October 7-8, 2013, Woods Hole Oceanographic Institution, available online at: http://tinyurl.com/lb7a82f

[2] D. L. McGuinness, P. Fox, B. Brodaric, and E. Kendall, The emerging field of semantic scientific knowledge integration. *Intelligent Systems, IEEE*, *24*(1), 25-26, 2009

[3] P. Fox and J. Hendler, Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science, in The Fourth Paradigm: Data-Intensive Scientific Discovery, Edited by Tony Hey, Stewart Tansley, and Kristin Tolle, Microsoft Research, pp 147–152, 2009

[4] E. Neumann, E., A Life Science Semantic Web: Are We There Yet?, Sci. STKE, Issue 283, May, 2005.

[5] P. Fox, D. L. McGuinness, D. Middleton, L. Cinquini, A. Darnell, J. Garcia, P. West, J. L. Benedict, and S. Solomon, Semantically-Enabled Large-Scale Science Data Repositories, In Proceedings of the International Semantic Web Conference, Lecture Notes in Computer Science 4273 (Cruz, I. et al., eds.), pp. 792–805, Springer, 2006

[6] J. Madin, S. Bowers, M. Schildhauer, S. Drivov, D. Pennington, and F. Villa, An ontology for describing and synthesizing ecological observation data. Ecology Information 2 (3), 279–296, 2007

[7] T. W. Narock, V. Yoon, J. Merka, and A. Szabo, The Semantic Web in Federated Information Systems: A Space Physics Case Study, Journal of Information Technology Theory and Application, Volume 11, Issue 3, Article 3, pp. 25-41, 2010.

[8] T. W. Narock and P. Fox, From Science to e-Science to Semantic e-Science: a Heliophysics Case Study, Computers & Geosciences, Volume 46, September, pp. 248-254, 2012

[9] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruichshank, M. Delderfield, I. Dunlop, M. Gamble, D. Michaelides, S. Owen, D. Newman, S. Sufi, and C. Goble, Why linked data is not enough for scientists, Future Generation Computer Systems, Volume 29, Issue 2, February 2013, pp. 599-611, 2013

[10] A. Leadbetter, T. Hamre, R. Lowry, Y. Lassoued, and D. Dunne, Ontologies and Ontology Extension for Marine Environmental Information Systems, Proceedings of the Workshop *Environmental Information Systems and Services - Infastructures and Platforms*, Edited by Arne J Berre, Dumitru Roman and Patrick Maue, (envip'2010), Bonn, Germany: CEUR Workshop Proceedings, 2010.

[11] R. Arko, C. Chandler, K. Stocks, S. Smith, P. Clark, A. Shepherd, C. Moore, and S. Beaulieu, Rolling Deck to Repository (R2R): Collaborative Development of Linked Data for Oceanographic Research, Geophys. Res. Abstr. Vol. 15; EGU2013-9564, 2013

[12] A. Gangemi, Ontology design patterns for semantic web content, in: Y. Gil, E. Motta, R. Benjamins, M. Musen (Eds.), 4th International Semantic Web Conference, Lecture Notes in Computer Science, ISWC 2005, vol. 3729, Springer, pp. 262–276, 2005

[13] K. Janowicz and P. Hitzler, The digital earth as knowledge engine, Semantic Web 3(3), pp. 213-221, 2012

[14] Shepherd, A., C. Chandler, R. Arko, Y. Chen, A. Krisnadhi, P. Hitzler, R. Groman, S. Rauch (2014), Semantic Entity Pairing for Improved Data Validation and Discovery, Geophys. Res. Abstr. Vol. 16; EGU2014-2476.

[15] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao, PROV-O: The PROV Ontology, W3C Recommendation 30 April 2013

[16] Chen, Y., A. Shepherd, C. Chandler, R. Arko, A. Leadbetter (2014), Ontology Based Vocabulary Matching Platform, Geophys. Res. Abstr. Vol. 16; EGU2014-12909.

[17] R. Isele, A. Jentzsch, and C. Bizer, Silk Server - Adding missing Links while consuming Linked Data, 1st International Workshop on Consuming Linked Data (COLD 2010), Shanghai, China, November, 2010.

[18] A. Schultz, A. Matteini, R. Isele, P. Mendes, C. Bizer, and C. Becker, LDIF - A Framework for Large-Scale Linked Data Integration, 21st International World Wide Web Conference (WWW2012) Developers Track. Lyon, France, April 2012.

[19] Parsons, M.A., Duerr, R., & Minster, J.-B. (2010). Data citation and peer-review. *Eos, Transactions, American Geophysical Union*,**91**(34), 297–298.

[20] T. W. Narock and P. Hitzler, Crowdsourcing Semantics for Big Data in Geoscience Applications, AAAI 2013 Fall Symposium Series, Semantics for Big Data, November 15-17, Arlington, Viginia, 2013