From data portal to knowledge portal: Leveraging semantic technologies to support interdisciplinary studies

Xiaogang Ma, Patrick West, John Erickson, Stephan Zednik, Yu Chen, Han Wang, Hao Zhong, Peter Fox

Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA
Email: {max7; westp; erickj4; zednis2; cheny18; wangh17; zhongh3; foxp}@rpi.edu

Abstract. Scientific research practices regularly adopt new technologies and platforms in an effort to increase information timeliness, sharing and discoverability. There are many initiatives related to open data, open code, open access, open collections, composing the topic of Open Science in academia. Being open has two levels of meanings. The first is to make the data, code, sample collections and publications, etc. freely accessible online. The other is the annotation and connection between those resources to establish the provenance information for reproducible scientific research. In this paper we present our work on a web portal for the Deep Carbon Observatory (DCO) community [1]. The DCO is a 10-year (2009-2019) initiative to intensify global attention and scientific effort in the burgeoning field of deep carbon science. Inspired by guiding questions such as "how much carbon does Earth contain?", "where is it?" and "what can deep carbon tell us about origins?" more than 1000 scientists across the world are actively participating in the DCO community. The DCO web portal is a research collaboration website developed to keep track of all researchers, organizations, instruments, field sites, and research outputs related to the DCO community. We intend for the DCO web portal to be a knowledge portal - adopting state-of-the-art semantic technologies to support various stages of the scientific process within and beyond the DCO community.

Keywords: Semantic Web; eScience; Knowledge Portal; Ontologies; Data Stewardship

1 A model of the science network

The context of our work is the Semantic Web, which is defined as an extension to the current Web by adding machine readable meanings and context to information on the Web. In this way, the Web is being transformed from a Web of Documents to a Web of Data [2]. Ontologies are an important way of capturing and representing machine readable meanings. An ontology is the formal specification of the shared conceptualization of a domain of study. In our work surrounding the DCO web portal, an initial part was the development of domain specific ontologies, and the integration of already existing ontologies. Our portal adapted the VIVO system as a platform for metadata management. The VIVO system itself already uses a list of ontologies to support academic information management. In our work we further extended the VIVO system by developing a DCO ontology and importing a few other ontologies such as the PROV Ontology [3] for provenance documentation and DCAT [4] to represent datasets and data catalogs. Table 1 lists the key ontologies and schemas used in the web portal.

Table 1 Ontologies and schemas used in the DCO web portal

Name	**Namespace URL**	**Prefix**
Dublin Core Metadata Element Set	http://purl.org/dc/elements/1.1/	dc
DCMI Metadata Terms	http://purl.org/dc/terms/	dct
VIVO Core	http://vivoweb.org/ontology/core#	vivo
VIVO Scientific Research Ontology	http://vivoweb.org/ontology/scientific-research#	scires
Data Catalog Vocabulary	http://www.w3.org/ns/dcat#	dcat
Bibliographic Ontology	http://purl.org/ontology/bibo/	bibo
Citation Counting and Context Characterization Ontology	http://purl.org/spar/c4o/	c4o
Citation Typing Ontology	http://purl.org/spar/cito/	cito
FRBR-Aligned Bibliographic Ontology	http://purl.org/spar/fabio/	fabio
Event Ontology	http://purl.org/NET/c4dm/event.owl#	event
Friend of a Friend	http://xmlns.com/foaf/0.1/	foaf
vCard Ontology	http://www.w3.org/2006/vcard/ns#	vcard
Geopolitical Ontology	http://aims.fao.org/aos/geopolitical.owl#	geo
Simple Knowledge Organization System	http://www.w3.org/2004/02/skos/core#	skos
DCO Ontology	http://info.deepcarbon.net/schema#	dco
PROV Ontology	http://www.w3.org/ns/prov#	prov

It should be noted that those ontologies are not separated from each other. Instead, they are integrated as a whole knowledge graph for representing the various agents, entities and activities in the DCO scientific community. Ontology reuse and intermapping built the relationships among the components in this knowledge graph. For example, bibo, c4o, cito and fabio ontologies represent the network of bibliographic and citation information among various types of publications, foaf represents the network of researchers and organizations, vivo and dco further extend the inter-connections among those components and other objects such as research topics, grants, projects, awards, and more. Provenance documentation leverages the W3C recommendation prov, which represents a high level framework. Classes and properties in other ontologies, such as dco, vivo and foaf, can be mapped as subclasses and subproperties of corresponding classes and properties in prov. Moreover, the knowledge graph can be extended according to real-world needs, especially dco, which is an ontology created and curated by ourselves for the DCO community.

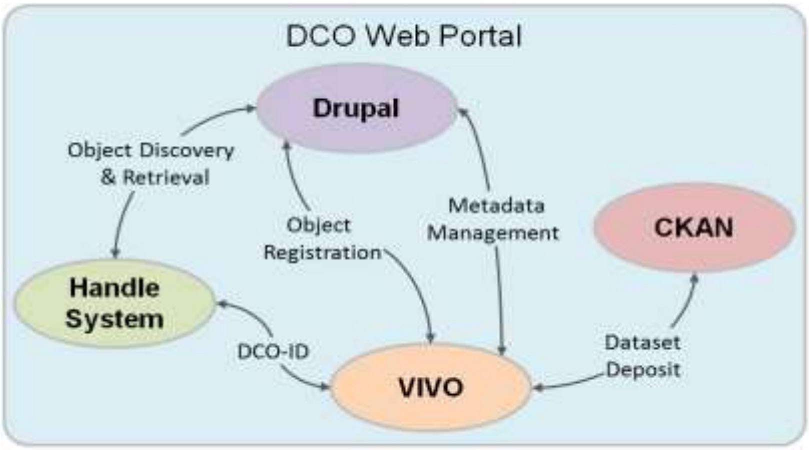


Fig. 1. A schematic view of the DCO web portal.

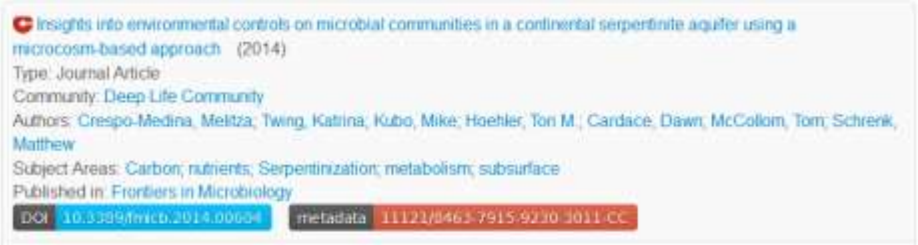
2 Annotation and linking to create semantics

With a knowledge graph as the core, the developed DCO web portal consists of four major parts: Drupal, a content management system used as the main front-end web portal where users can register, discover and retrieve various types of objects; Handle System, which is used to assign a persistent and unique identifier to the objects, known as a DCO-ID; VIVO, the main knowledge store; and CKAN, used for the storage and archiving of datasets and other media. (Figure 1). Object registration, discovery and retrieval can all be facilitated through the use of the DCO-ID, similar to what the Digital Object Identifier (DOI) does for publications.

The functionalities of the web portal enable an individual researcher to record almost all the components in the life cycle of their research, from funding application, instrument deployment, field work planning, data collection, data analysis, meeting records, publication archival, to project reporting, and more. The portal also allows researchers who share a common research interest to find, communicate with and collaborate on research through virtual groups. All instance objects can be annotated with a list of properties from the corresponding ontologies and can be linked directly or indirectly to other objects. For example, a journal paper can be tagged with a few keywords as its topics. One or more of those keywords may also be used to represent the research interests of a researcher, who might find that paper of interest by searching the keywords, and from the keywords the researcher may in turn find other publications or researchers within the same domain. Such annotation and interconnections among instance objects provide a more detailed network about the real world situation and can expand our understanding of the research to an extent that cannot be reached by only reading the conventional publications.

3 Identification and persistence of DCO resources

The DCO-ID provides a persistent and unique identifier to all resources in the DCO web portal. The DCO-ID is similar to the DOI for publications, but it extends the scope to many more types of objects, including publications, people, organizations, instruments, datasets, sample collections, keywords, conferences, etc. The environment of the Web may evolve in the future and the web addresses of the portal and the various objects registered in it may change. With the DCO-ID, even after 10 or 100 years, one can still find the associated web address of that object and retrieve the information needed. In this way we can keep a persistent and stable legacy for the activities and outputs of the DCO community. Fig. 2 shows a DCO publication records, which has both a DOI and a DCO-ID (shown as a code in the 'metadata' bar). The DCO-ID allows users to retrieve more domain specific annotations from the metadata of the publication in the portal. The records of community, authors, subject areas and journal shown in Fig. 2 are all hyperlinks and they all have their own DCO-IDs.

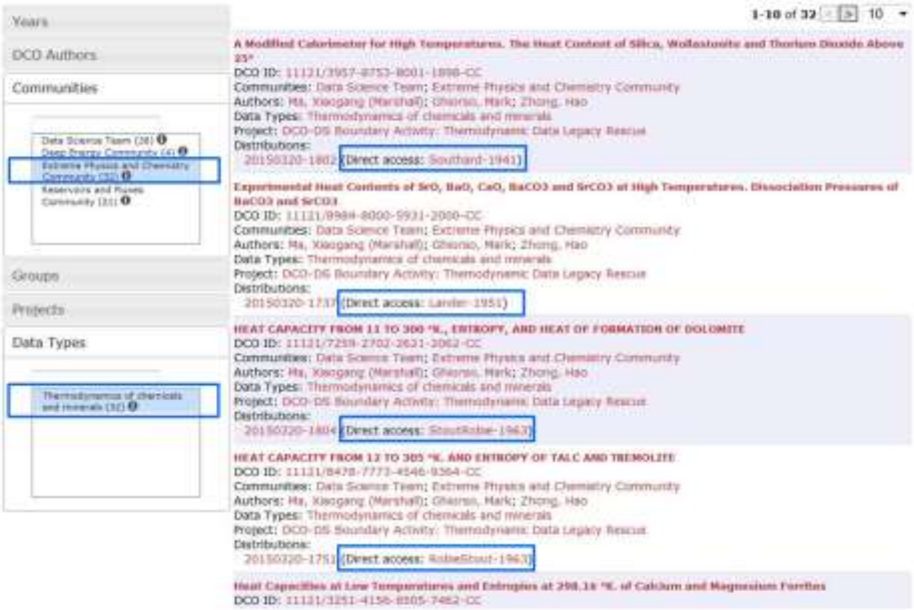


Insights into environmental controls on microbial communities in a continental serpentinite aquifer using a microcosm-based approach (2014)
Type: Journal Article
Community: Deep Life Community
Authors: Crespo-Medina, Melitza; Twing, Katrina; Kubo, Mike; Hoehler, Tori M.; Cardace, Dawn; McCollom, Tom; Schrenk, Matthew
Subject Areas: Carbon; nutrients; Serpentinization; metabolism; subsurface
Published in: Frontiers in Microbiology
DOI 10.3389/fmicb.2014.00604 metadata 11121/8463-7915-9230-3011-CC

Fig. 2 A record in the DCO publication browser.

4 State-of-the-art data stewardship

Data stewardship has a two-fold meaning: data management and data service. Semantic technologies can leverage both parts. The above sections focus more on the data management side. In our work we also made innovative progress on the data service side. We are working together with other organizations to advance discovery and usability of science data as well as other resources. One recent collaboration is with the output of the Data Type Registry (DTR) working group [5] of the Research Data Alliance. Each DTR is a self-contained portal for data type registration and curation. There are some common basic types, which are called 'primitives' and will be registered and managed by a central data type registry. This shows a two level hierarchy of a DTR, one is a list of primitives and the other is the specific data types defined within a DTR. This two-level hierarchy initiated our extension to the DCO ontology. In our work, the basic data types are classes in the DCO ontology, and the specific data types are at the instance level, i.e., they are all instances of a newly created class "dco:DataType" and are part of our knowledge graph but accessible outside of DCO. Fig. 3 shows the DCO dataset browser, in which the data type is used a facet that can help users to find dataset of interest.


Years
DCO Authors
Communities
Data Science Team (28)
Deep Energy Community (4)
Extreme Physics and Chemistry Community (32)
Reservoirs and Fluxes Community (21)
Groups
Projects
Data Types
Thermodynamics of chemicals and minerals (32)
1-10 of 32
10

A Modified Calorimeter for High Temperatures. The Heat Content of Silica, Wollastonite and Thorium Dioxide Above 25°
DCO ID: 11121/3957-8753-8001-1898-CC
Communities: Data Science Team; Extreme Physics and Chemistry Community
Authors: Ma, Xiaogang (Marshall); Ghiorso, Mark; Zhong, Hao
Data Types: Thermodynamics of chemicals and minerals
Project: DCO-DS Boundary Activity: Thermodynamic Data Legacy Rescue
Distributions:
20150320-1802 (Direct access: Southard-1941)

Experimental Heat Contents of SrO, BaO, CaO, BaCO3 and SrCO3 at High Temperatures. Dissociation Pressures of BaCO3 and SrCO3
DCO ID: 11121/8984-8000-5931-2000-CC
Communities: Data Science Team; Extreme Physics and Chemistry Community
Authors: Ma, Xiaogang (Marshall); Ghiorso, Mark; Zhong, Hao
Data Types: Thermodynamics of chemicals and minerals
Project: DCO-DS Boundary Activity: Thermodynamic Data Legacy Rescue
Distributions:
20150320-1737 (Direct access: Lander-1951)

HEAT CAPACITY FROM 11 TO 300 °K., ENTROPY, AND HEAT OF FORMATION OF DOLOMITE
DCO ID: 11121/7259-2702-2631-3062-CC
Communities: Data Science Team; Extreme Physics and Chemistry Community
Authors: Ma, Xiaogang (Marshall); Ghiorso, Mark; Zhong, Hao
Data Types: Thermodynamics of chemicals and minerals
Project: DCO-DS Boundary Activity: Thermodynamic Data Legacy Rescue
Distributions:
20150320-1804 (Direct access: StoutRobie-1963)

HEAT CAPACITY FROM 12 TO 305 °K. AND ENTROPY OF TALC AND TREMOLITE
DCO ID: 11121/8478-7773-4546-9364-CC
Communities: Data Science Team; Extreme Physics and Chemistry Community
Authors: Ma, Xiaogang (Marshall); Ghiorso, Mark; Zhong, Hao
Data Types: Thermodynamics of chemicals and minerals
Project: DCO-DS Boundary Activity: Thermodynamic Data Legacy Rescue
Distributions:
20150320-1751 (Direct access: RobieStout-1963)

Heat Capacities at Low Temperatures and Entropies at 298.16 °K. of Calcium and Magnesium Ferrites
DCO ID: 11121/3251-4156-8805-7462-CC

Fig. 3 Using data type as a facet to retrieve datasets of interest.

Besides dataset curation, we also utilize the latest progress on data citation and sample collection curation. For example, we created a class 'dco:GeoSample' in the DCO ontology which refers to the global initiative International GeoSample Number [6] for metadata used to annotate geological samples. We also use the metadata schema DataCite [7] for citation properties of registered datasets in the DCO web portal.

5 Concluding remarks

Our aim for the DCO web portal is to create more than just a data portal, but a knowledge portal. By using semantic technologies and leveraging state-of-the-art methods in data stewardship we built a web portal for the DCO community to support various aspects of their research. The information collected in the portal, from both the DCO community and extramural data resources, is stored in ways that both humans and computers can read and understand. A key feature of our portal, as enabled by the Semantic Web, is the linkage among various registered objects and the flexible ways to present them. With linked data we are able to create more and better collaborations, find like-minded individuals working on common projects, add data that can be useful to others, discover tools that can be used to visualize data in new ways, and make it easier to discover, access, understand and use the data.

References

1. https://deepcarbon.net/
2. Berners-Lee, T., Hendler, J., Lassila, O., 2001. The Semantic Web. Scientific American 284 (5), 34-43.
3. Lebo, T., Sahoo, S., McGuinness, D., 2013. PROV-O: The PROV Ontology. Accessible at: http://www.w3.org/TR/prov-o/
4. Maali, F., Erickson, J., 2014. Data Catalog Vocabulary (DCAT). Accessible at: http://www.w3.org/TR/vocab-dcat/
5. https://rd-alliance.org/groups/data-type-registries-wg.html
6. http://www.igsn.org/
7. https://www.datacite.org/node