frontiers in Earth Science

ORIGINAL RESEARCH
published: 15 May 2017
doi: 10.3389/feart.2017.00036

Weaving a Knowledge Network for Deep Carbon Science

Xiaogang Ma[1,2], Patrick West[2], Stephan Zednik[2], John Erickson[2], Ahmed Eleish[2], Yu Chen[2], Han Wang[2], Hao Zhong[2] and Peter Fox[2]*

[1] Department of Computer Science, University of Idaho, Moscow, ID, USA, [2] Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA

Geoscience researchers are increasingly dependent on informatics and the Web to conduct their research. Geoscience is one of the first domains that take lead in initiatives such as open data, open code, open access, and open collections, which comprise key topics of Open Science in academia. The meaning of being open can be understood at two levels. The lower level is to make data, code, sample collections, and publications, etc., freely accessible online and allow reuse, modification, and sharing. The higher level is the annotation and connection between those resources to establish a network for collaborative scientific research. In the data science component of the Deep Carbon Observatory (DCO), we have leveraged state-of-the-art information technologies and existing online resources to deploy a web portal for the over 1,000 researchers in the DCO community. An initial aim of the portal is to keep track of all research and outputs related to the DCO community. Further, we intend for the portal to establish a knowledge network, which supports various stages of an open scientific process within and beyond the DCO community. Annotation and linking are the key characteristics of the knowledge network. Not only are key assets, including DCO data and methods, published in an open and inter-linked fashion, but the people, organizations, groups, grants, projects, samples, field sites, instruments, software programs, activities, meetings, etc., are recorded and connected to each other through relationships based on well-defined, formal conceptual models. The network promotes collaboration among DCO participants, improves the openness and reproducibility of carbon-related research, facilitates accreditation to resource contributors, and eventually stimulates new ideas and findings in deep carbon-related studies.

Keywords: data stewardship, knowledge network, eScience, semantic web, ontologies

OPEN ACCESS

Edited by:
Donato Giovannelli, Earth-Life Science Institute – Tokyo Institute of Technology, Japan

Reviewed by:
Alessandro Sarretta, Consiglio Nazionale delle Ricerche – Istituto di Scienze Marine, Italy
Christian Schröder, University of Stirling, UK
Holly M. Bik, University of California, Riverside, USA

****Correspondence:***
Xiaogang Ma
max@uidaho.edu
xgmachina@gmail.com

Specialty section:
This article was submitted to Geochemistry, a section of the journal Frontiers in Earth Science

Received: *13 January 2017*
Accepted: *28 April 2017*
Published: *15 May 2017*

Citation:
Ma X, West P, Zednik S, Erickson J, Eleish A, Chen Y, Wang H, Zhong H and Fox P (2017) Weaving a Knowledge Network for Deep Carbon Science. Front. Earth Sci. 5:36. doi: 10.3389/feart.2017.00036

INTRODUCTION

Recent advances in cyberinfrastructure facilitate the culture of open science (Nosek et al., 2015) and also provide a space for conducting scientific work in a more efficient way. The geoscience community has taken an active role in the discussion and efforts on open access publication (Harnad and Brody, 2004), open samples (Lehnert et al., 2006), open source software (Hey and Payne, 2015), and open data (Glaves, 2017). Geoscience researchers are increasingly dependent on information technologies and the World Wide Web to conduct and communicate their research. The keyword "open" in those open science efforts does not mean to publish individual works or resources as separated fragments. Instead, those resources can be categorized, annotated, and

connected to each other, and thus form a knowledge network (Ma et al., 2014b). In such a network, each node has its resource type and is described with detailed and meaningful information, such as a rock sample registered at a museum (Devaraju et al., 2016). There are also various types of relationships connecting among nodes. For instance, there could be a relationship "registrant" which connects a rock sample to a researcher. In turn, there could be another relationship "author of" which connects the researcher to a number of publications (**Figure 1**). Such an open knowledge network has a lot of potential uses in resource discovery and access, program administration, research collaboration, scientometrics, research trend analysis, and more. To build and implement such a network, however, needs cross-disciplinary collaboration to identify the information to be covered as well as state-of-the-art methodologies and technologies to realize the network in an operable platform. In a recent research program called Deep Carbon Observatory (DCO), under the Data Science activity, we successfully carried out a study in that direction and put it into practical use for the DCO community.

DCO is a 10-year (2009–2019) global scientific initiative focusing on the study of carbon in deep Earth. More than 1,000 researchers across the world have been participated in this initiative, and the research is organized into four science communities—Deep Energy, Deep Life, Extreme Physics, and Chemistry and Reservoirs and Fluxes. Funded activities include computational model development, new instrument design, novel technology application, exploratory research, and fieldwork, scientific conferences, and early career summer school/meetings. Facing the opportunities enabled by the cyberinfrastructure and open science, the DCO community has been seeking new ways of data management and data analytics through various DCO data science activities led by the team at Rensselaer Polytechnic Institute. A most recent achievement of those activities is a knowledge network, which is now in its practical version. The platform for the implementation of this network, called the Deep Carbon Virtual Observatory (DCvO), has the following characteristics and functions: a schema categorizing various concepts in DCO scientific works, a capability to define, and describe provenance information such as key entities, agents, and activities in a workflow, a repository for storing research datasets and the metadata annotating them, tools for scientific communication, and collaborations, and an integrated portal with friendly user interface for managing all those resources and applications. DCvO sets up an environment to facilitate open science within the DCO Community (Fox, 2015).

DCvO provides a means to collaborate, seek out and access education, and research materials such as publications and datasets, view community activities, view DCO field sites, learn about funded projects, and more. The intent of the DCvO, just like DCO itself, is to stimulate new ways to conduct and share deep carbon-related research. Underlying DCvO is the application of ideas, practices, and technologies from the fields of information science, data management, data analytics, computer science, and physical sciences using contemporary cyberinfrastructure and information technologies. DCvO promotes collaboration among DCO participants, improves the openness and reproducibility of carbon-related research, facilitates accreditation to resource contributors, and eventually stimulate new ideas and findings in deep carbon-related studies. The rest of the paper will introduce the methods and technological components applied to construct the DCO knowledge network, and features of a few realized modules in DCvO.

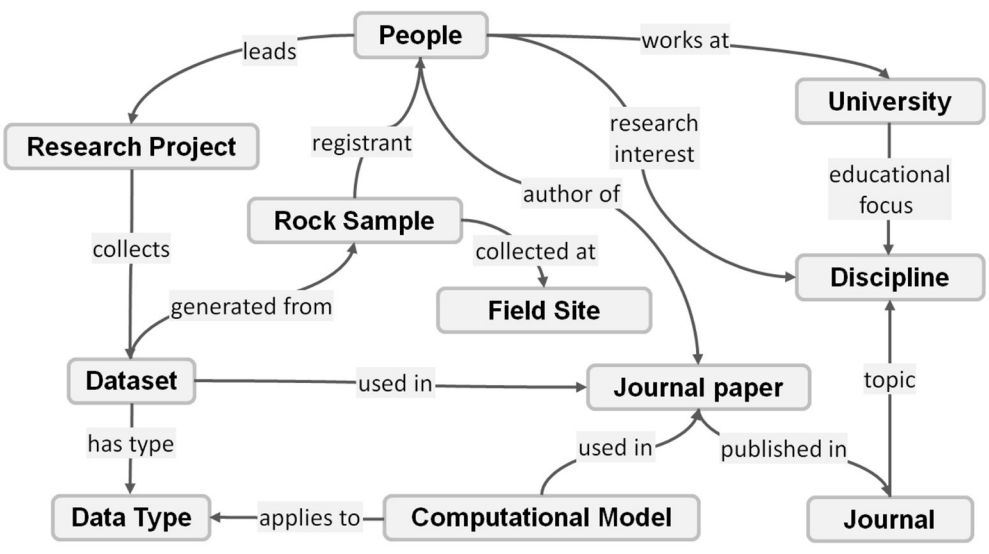

FIGURE 1 | A fragment of a network showing part of the agents, entities and activities involved in scientific research.

METHODS AND TECHNOLOGIES

Add Structures and Meanings to Content

To make the output of our work easily accessible by geoscientists and address the features of open science, we conducted our work within the context of the World Wide Web and used state-of-the-art web technologies. More specifically, the context of our work is the Semantic Web, which is defined as an extension to the current Web by adding machine readable structures, meanings, and context to information on the Web (Berners-Lee et al., 2001). The Web is now in the transition from a Web of Documents to a Web of Data because of the embedded structures and meanings that did not exist before. Nevertheless, to add structure and meaning to the information on the Web, formal specifications of concepts and the interrelationships among concepts are needed. In the Semantic Web such formal specifications are called ontologies. Each ontology is the formal specification of the shared conceptualization (Gruber, 1995) of a domain of study. In an inter-disciplinary context there thus could be a large number of ontologies. While those ontologies each address a certain topic, there could be interrelationships among them, which are the key to weave a bigger knowledge graph. In our work for the DCO science and the DCvO, a primary work is to recognize domain specific ontologies to be reused or developed, and the integration of all those ontologies into an umbrella ontology called the DCO ontology (http://info.deepcarbon.net/schema). The core of DCO ontology is the VIVO ontology (Mitchell et al., 2011), which reuses and extends a list of ontologies to support information management in the academia. In DCO ontology we further extended the VIVO ontologies by adding concepts and relationships recognized from the DCO science needs (Ma et al., 2015) and also by reusing several other ontologies (**Table 1**), including the PROV Ontology (Lebo et al., 2013) for provenance documentation and DCAT (Maali and Erickson, 2014) to represent data catalogs.

As noted above, those sub-ontologies within the DCO ontology are not isolated from each other but are inter-connected as a knowledge network for representing the objects in the scientific workflow (cf. PROV-O core model) as well as their interrelationships for the DCO scientific community. The reuse and adaptation of those ontologies is driven by needs in the DCO science community and we applied a use case-driven approach (Fox and McGuinness, 2008) to analyze the needs. Subsequently, we identified and collected the ontologies to be reused, or created new concepts and relationships as an extension to existing ontologies. The inter-mappings among those ontologies (**Table 1**) set up the foundation for the DCO science knowledge network. For example, four ontologies bibo, c4o, cito, and fabio were used to record individual records and the network of bibliographic information, foaf was used for individuals and the network of researchers and organizations. Then vivo and dco ontologies were used to extend the inter-connections among researchers and publications, as well as other objects such as projects, grants, keywords, funding awards, and more. Our work of provenance records in DCvO leveraged the W3C standard prov (https://www.w3.org/TR/prov-o/), which represents a high level framework. Components in a few domain ontologies, such as dco, foaf, and vivo can be mapped as subclasses or sub-properties of corresponding parts in prov (Ma et al., 2014c).

TABLE 1 | Ontologies and schemas used in DCvO.

Name	Namespace URL	Prefix
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
Data catalog vocabulary	http://www.w3.org/ns/dcat#	dcat
DCMI metadata terms	http://purl.org/dc/terms/	dct
DCO schema	http://info.deepcarbon.net/schema#	dco
Dublin core metadata element set	http://purl.org/dc/elements/1.1/	dc
Event ontology	http://purl.org/NET/c4dm/event.owl#	event
FRBR-aligned bibliographic ontology	http://purl.org/spar/fabio/	fabio
Friend of a friend	http://xmlns.com/foaf/0.1/	foaf
Geopolitical ontology	http://aims.fao.org/aos/geopolitical.owl#	geo
PROV ontology	http://www.w3.org/ns/prov#	prov
Simple knowledge organization system	http://www.w3.org/2004/02/skos/core#	skos
vCard ontology	http://www.w3.org/2006/vcard/ns#	vcard
VIVO core	http://vivoweb.org/ontology/core#	vivo
VIVO scientific research ontology	http://vivoweb.org/ontology/scientific-research#	scires

A Technological Framework for DCvO

The knowledge graph realized by the DCO ontology set up a framework for filling in detailed records from the DCO community. Through the knowledge graph those records will have structured description and are interconnected with meaningful relationships, which as whole we referred to as the DCO knowledge network. To weave that knowledge network and implement it in DCvO, a technological framework was developed (**Figure 2**), which consisted of four major parts: Drupal (https://www.drupal.org) was used to develop the main front-end portal where users can contribute and access various types of information; The Global Handle System (https://www.handle.net) was used to assign a unique identifier called DCO-ID, to the records in the knowledge network; VIVO (http://vivoweb.org) was used as the main knowledge store for archiving the ontologies and the detailed records; and CKAN (http://ckan.org) was used for the storage of datasets and other media files.

DCvO enables individual researchers or collaborative groups to record and connect most objects in the procedure their research life cycle, from funding applications, instrument proposal, field investigation, data processing, experiment documentation, to publication archival, and project report, and more. In the virtual environment of DCvO, researchers can

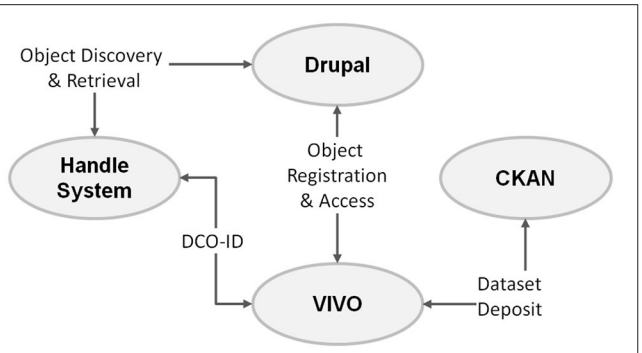

FIGURE 2 | A schematic view of the DCvO technological components.

register details about their previous activities and publications as well as current research interests. Such information will enable researchers sharing similar interests to discover and communicate with each other and propose future works. In DCvO, each DCO-ID represents an object and corresponds to a Web address (URL). Once a researcher knows the DCO-ID of a certain object, he can easily access more detailed information in the DCvO by resolving a DCO-ID in a Web browser. All instance object records in the knowledge network can be annotated with selected subjects from one or a few common ontologies and vocabularies. In this way, those objects are linked directly or indirectly to each other. For instance, a researcher tags a dataset uploaded by him with a few labels as its keywords. Some of those labels are used by another researcher as his research interests. For the latter researcher, he can easily find that the dataset can be of interest by using those labels in data search. Similarly, those labels can also be used as keywords for other resources such as funded projects, journal papers and research institutions and will make it easier for the researcher to find those resources. Such features of annotated and interconnected resources in DCvO can help expand researchers' understanding of his research and promote efficient conduction.

IMPLEMENTATION AND RESULTS

For end users of the knowledge network, either from the DCO community or the general public, the most familiar site is the DCO community portal (**Figure 3**) which was developed based on Drupal. As a member of the community, i.e., registered, one can submit ad hoc content like news and information of events to the community website, and via e the menu bar item "Data Portal" can explore and use resources associated with the knowledge network. All content in the knowledge network is open to the public, including the DCO community. However, in order to be able to add and/or update records in the knowledge network one needs to be a registered member, which can be done through a member registration process. The following sections introduce a few of the functions enabled by using the knowledge network.

Semantically Enriched Faceted Browsers

To enhance the publication search capability in the DCvO, faceted browsing (Ellis and Vasconcelos, 1999) has been utilized. This interface is accessible on the DCO website (**Figure 4**) and undergoes regular enhancements. For example, in the last year more facets were added (including DCO Community, author, and year) to expand selection choices, and more information about selections such as publication type and related DCO community is displayed in the result set, and more features available to enhance the users' ability to search for and retrieve publications. Along the left of the browser shown in **Figure 4** are expandable facets, which can be used to create a specific search of the database. Each facet represents a certain type of object in DCO. The appearance of linked entries among those facets represent a view of the DCO knowledge network and that such linked content can appear on many pages or in browsers under different contexts (e.g., publications, field sites, etc.). For example, if a user wants to search by authors, he can expand the author facet. Or, the user can search using keywords after expanding that facet. The search box at the top allows free text search and queries all information. In another example, if a user wants to find all the publications related to "molecular biology," he can just type that word in the search box. Or if he wants to see all publications authored by Robert Hazen, he can just type the last name "Hazen" in the search box and hit the search button.

The search and retrieve mechanism of the faceted browser greatly speeds up information discovery and access. The DCO knowledge network is recorded in a Semantic Web triple store where all DCO metadata are stored as linked data. That store can handle complex queries for discovering concepts and the relationships among them, retrieving complex information and relationships, and more and rendered on almost all of the web pages on the DCO website. The disadvantage of the particular triple store in use, in most cases, is that as more and more information is stored, the slower text-based searches become. Although the technology has taken great leaps forward in performance and scalability over the last couple of years, it is still relatively slow. Triple stores are best suited to relationship queries and not free-text.

As a result, the improved publications browser information is ingested from the triple store into an inverted index, using an open source product called ElasticSearch (https://www.elastic.co). Using an inverted index approach (Seo et al., 2003) allows for searching over a great amount of text including keywords, abstracts, descriptions, author names, etc., very quickly. Thus, in the publications faceted search, information is displayed much faster than a browser that queries the triple store directly. Using the same technologies we also developed faceted browsers for other objects, such as datasets, people, grants, field studies, and more.

Scientific Data Types for DCO Data

DCO data come in all shapes and forms. Tables, for example, contain data but often lack context, or mix data with metadata. Such context might include the meanings of quantity names and units, acronyms, or community jargon, or the inter-relatedness of data columns or rows. To address such issues, we have

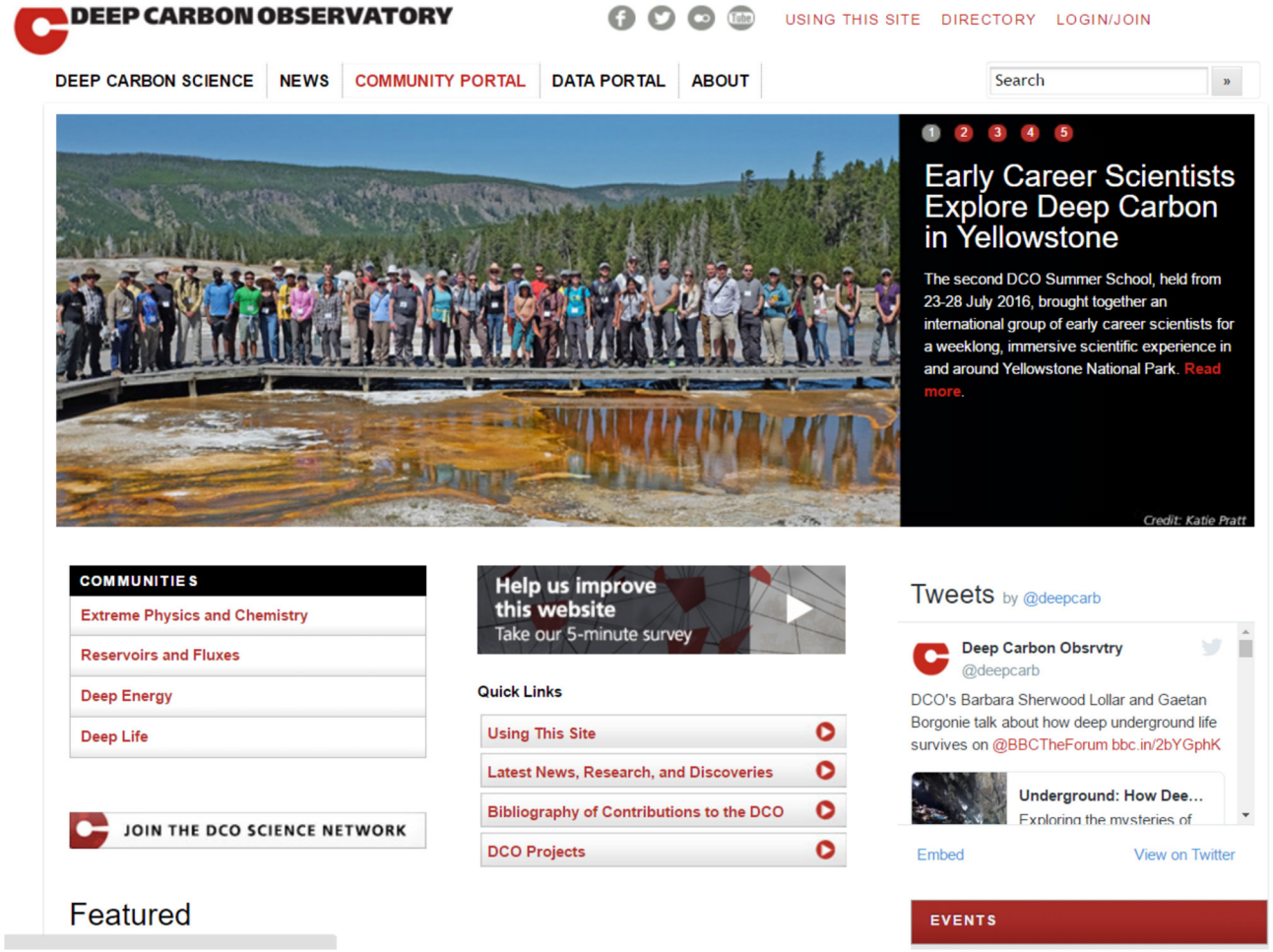

FIGURE 3 | A screenshot of the DCO community website (https://deepcarbon.net).

been working on ways to make disparate datasets accessible to scientists in diverse scientific disciplines by focusing on scientific context (Ma et al., 2016).

Traditional data types typically are limited to the numerical type of the datum, such as integer, float, array, char (character), or string. For example, a researcher might receive a table of numbers from a colleague, the title of which includes the word "Thermodynamics"; data that are relevant to their research. Beyond this, the table's data may be represented only by column headers, typically acronyms, one hopes well-known in the particular scientific domain. Moreover, any description of the relationships between the table's columns is not obvious, let alone explicit. However, many researchers think of categories of data, i.e., higher level ways of describing data they generate. For example: Volcanic Gas Composition. To enable such science context in computer-enabled data environments, the notion of "data type" must be extended so that a given data type represents a scientifically useful description of what the datasets associated with the category name actually represents. The ability to specify data types in this fashion enables researchers to better understand the meaning and ultimately usefulness or relevance, of datasets in a given scientific context.

The large number of DCO datasets currently registered by many scientists in DCO's four communities include a wide variety of formats and quantities with associated metadata about basic data types spanning Earth sciences, biological sciences, and beyond. The metadata (collected when datasets are registered with DCO) enables researchers to find and access DCO datasets via the Dataset Browser. A scientist might have a very specific request in mind though, such as "I need thermodynamic data from the DCO Extreme Physics and Chemistry Community that includes Mineral Name and Molecular Weight." Without proper metadata annotations, addressing this specific, but likely common, request is difficult, not only among DCO scientists but across scientific domains. Until recently, no widely agreed upon approach had been taken to address requests with scientific context, and researchers had to resort to other means of finding/assessing datasets.

Such broad issues in data management are the focus of initiatives such as the Research Data Alliance (RDA,

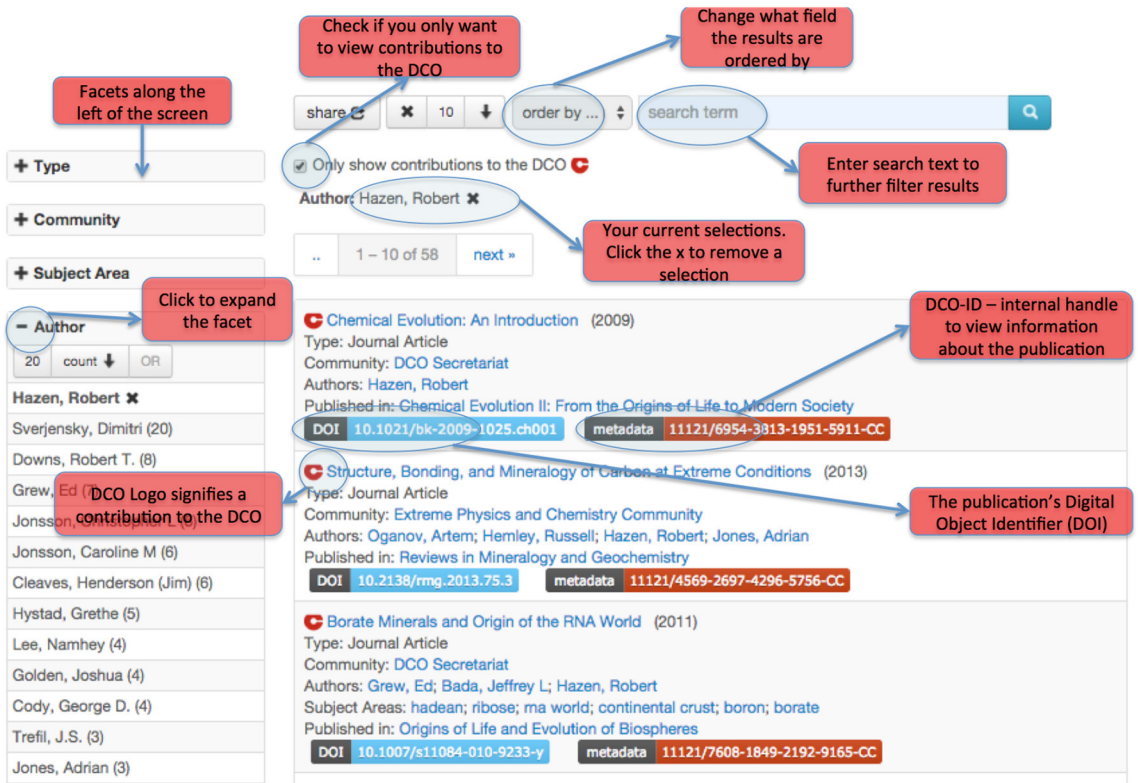

FIGURE 4 | A faceted browser for DCO publications (https://info.deepcarbon.net/vivo/publications).

https://www.rd-alliance.org). RDA is an international effort whose mission is to "... build the social and technical bridges that enable open sharing of data across technologies, disciplines, and countries." One of the bridge-building efforts to improve data sharing and data use in science communities involves a framework for developing "scientific data types." In 2015, we leveraged some funding from the National Science Foundation, via the RDA to make the DCvO one of the first platforms adopting two key RDA recommendations that greatly improved the modeling of scientific data types. The RDA deliverables adopted by DCO are the Data Type Registry (DTR) and Persistent Identifier Information Types (PIT). The first addresses a core interoperability problem among data management systems: the ability to parse, understand, and potentially reuse data retrieved from others. Scientific data types are visible to users. The second addresses the essential types of information associated with persistent identifiers [e.g., identifiers for people such as ORCID (https://orcid.org) are of a different "type" than identifiers for publications such as DOIs]. Permanent identifier types are largely invisible to users.

The curation and reuse of registered datasets within DCvO was well suited for testing deployments of RDA DTR and PIT because it helps address the challenges described in the above example of searching for thermodynamic datasets of interest, and provided valuable experience for other science communities who face the same issues.

In its implementation of RDA DTR and PIT, we first made updates to the DCO ontology, the backbone of DCvO, to incorporate concepts of data type and associated attributes. We collected data type instances from the DCO community and used them to annotate some initial datasets currently registered with DCvO. Results from this work are evident in the faceted DCO dataset browser and data type browser (**Figure 5**).

The above example, a researcher looking for thermodynamic data that includes Mineral Name and Molecular Weight, can first look for a corresponding data type using the data type browser. In that browser he can search Mineral Name and Molecular Weight in the facet window for Parameters and through which he can locate a data type, such as Thermodynamics of Chemicals and Minerals. Once the researcher finds that information, he can go to the dataset browser and retrieve all relevant datasets using that known data type. The researcher can also use the DCO Communities facet to restrict results to a subset, i.e., those generated by the DCO Extreme Physics and Chemistry Community. Using and expanding the registered (science context) data types, we foresee some future innovation, such as recommending datasets to a user based on his research interests and recommending tools for data analysis for specific data types. Such efforts will significantly facilitate work on data curation and promote the sharing and usability of deposited data. DCvO is also open for DCO science communities to add or suggest new data types for their datasets.

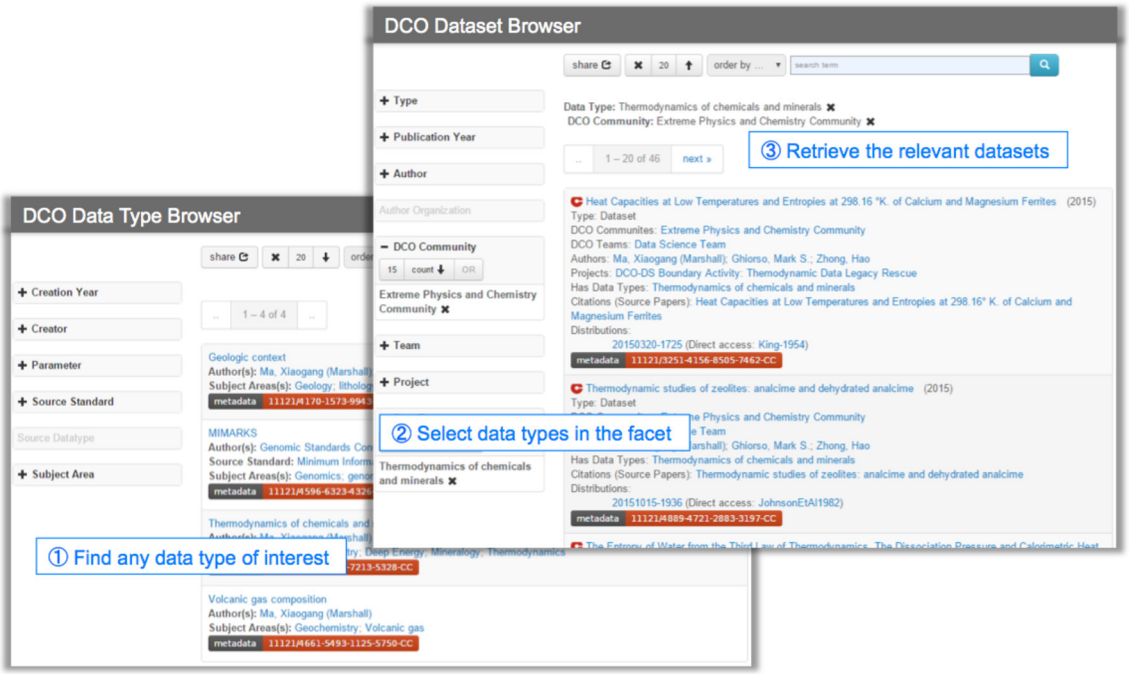

FIGURE 5 | Using DCO data type browser and DCO dataset browser to find datasets of interest.

Thermodynamic Data Rescue

A huge number of legacy datasets are contained in geoscience literature, often in the form of tables, and figures. Extracting, organizing, and reusing these datasets is valuable for many within the Earth and planetary science community. To explore methods and techniques for data rescue and management, we and Prof. Mark Ghiorso, and Extreme Physics and Chemistry community member identified thermodynamic datasets related to carbon as a proof of principle analysis, with a focus on records about the enthalpy and entropy of chemicals. The team developed a semi-automatic method for doing this. First, Ghiorso collected papers of focused themes in the fields of mineralogy, geochemistry, and petrology. We then extracted, reviewed, and registered the datasets via DCvO following the guidelines listed in the DCO data policy (https://deepcarbon.net/page/dco-open-access-and-data-policies). Most of those collected papers were published before the 1980s. Although provided in PDF format, their contents were scanned from printed copies and were saved as images. The workflow ensures extracted records are correct, well-organized and are saved in known formats. The resulting datasets are published and made discoverable through the DCO dataset browser (**Figure 6**).

To date, the team has dealt with three main types of datasets: (1) heat content or enthalpy data determined for a given compound as a function of temperature using high-temperature calorimetry, (2) heat content or enthalpy data determined for a given compound as a function of temperature using adiabatic calorimetry, and (3) direct determination of heat capacity of a compound as a function of temperature using differential scanning calorimetry. We have collected publications with additional thermodynamic sources, and an effort to rescue more datasets is ongoing. This will lead to a comprehensive characterization of the thermodynamics of carbon and carbon-related materials.

During the work, the team preserved individual datasets from various "frozen" and "dark" places as an open and stable data legacy via DCvO. Besides data registration and deposit (**Figure 2**), DCvO retains essential metadata for data discovery, use, and citation, as well as connections from the datasets to their original sources. The team archived each paper as a distinct data source, and collectively these data sources are searchable in DCvO. The "inter-connection" feature of the DCO knowledge network provides a mechanism for connecting rescued datasets beyond their individual data sources, to research domains, DCO Communities, and more; all of which make data discovery and retrieval more effective.

Leveraging Existing Resources to Create a Deep Carbon Data Legacy

The data portal in DCvO was designed to be a place for registering and archiving datasets that are generated from the DCO community as well as the global geoscience community. The "open access" feature of the portal has a meaning of 2-fold. First, the datasets on the portal are open for global access following the DCO data policy. Second, the portal is open for registering metadata of datasets that are stored in other data repositories, such as EarthChem, Pangaea, and National Centers for Environmental Information (NCEI), and others. The

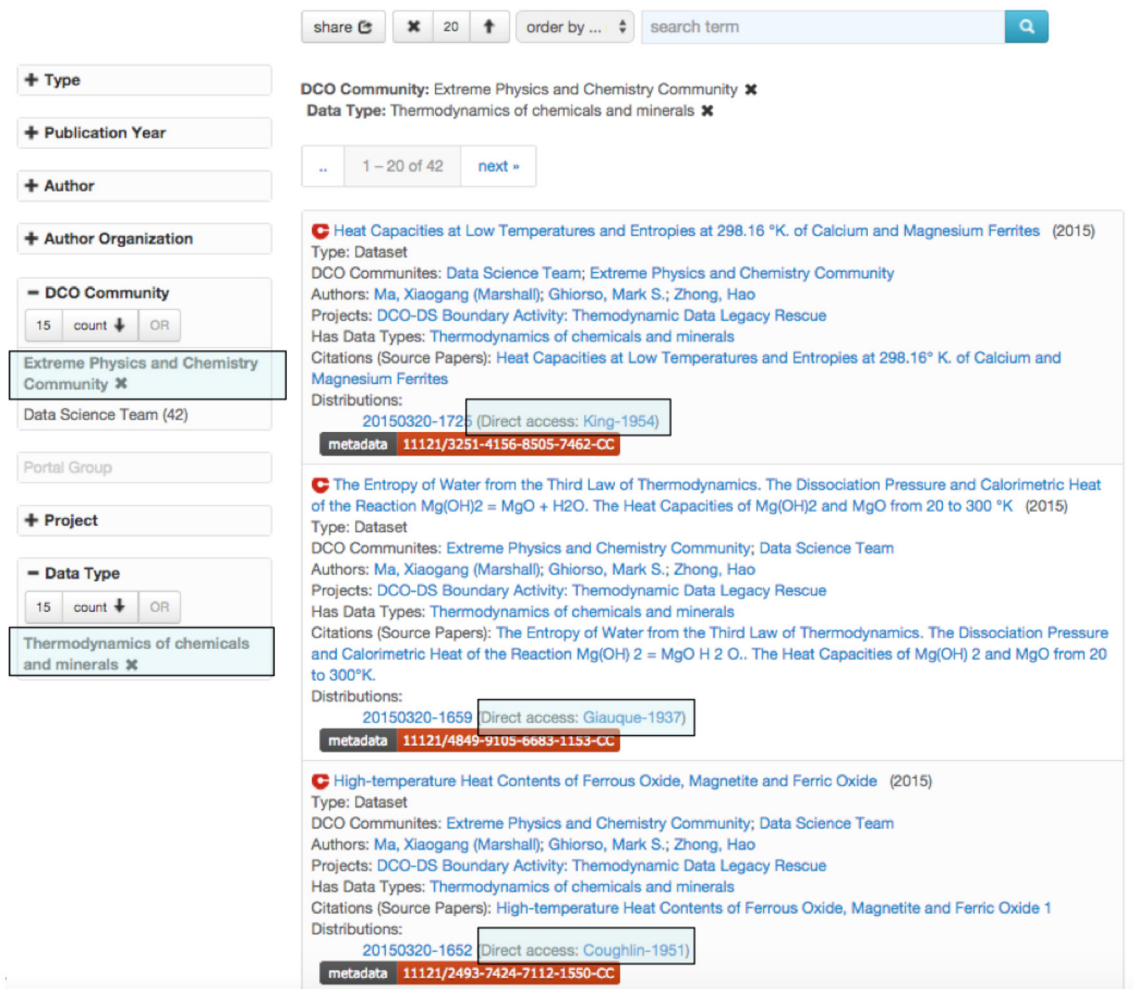

FIGURE 6 | Accessing rescued thermodynamic datasets in the DCO dataset browser. Rectangles indicate key selection concepts and returned results (the datasets themselves).

registration does not retrieve and archive a copy of the dataset in DCvO. Instead, it only deals with metadata. By using the original identifier (e.g., DOI) of the dataset, DCvO generates a redirection to the dataset in its original repository. The strategy of having DCO data collections already stored in existing and sustained community repositories increases the likelihood that these data legacies will continue to be available and valued well after the end of the DCO decade in 2019.

A recent example is the Legacy Russian Volcanic and Hydrothermal Gas Data. The datasets include gas samples from 20 publications, which were previously inaccessible to the non-Russian speaking community. In early 2015, Prof. Tobias Fischer (DCO DECADE program leader), together with expert Russian gas geochemist Yuri Taran (Universidad Autonoma de Mexico), Elena Kalacheva (IVS, Kamchatka), and Nicole Thomas (UNM), translated and compiled those datasets and archived them in EarthChem (See: http://earthchem.org/featured/fischer). As those datasets are of interest to the DCO Reservoirs & Fluxes community, their metadata were also registered in DCvO. The EarthChem DOIs of those datasets allow users navigate from DCvO pages to the dataset download links on EarthChem (**Figure 7**).

On the data portal of DCvO, each registered dataset has a DCO-ID. The portal also provides several other metadata items that can be used to enrich the description of the dataset, such as associated DCO community, subject area, geographic focus, data type, and more. The enriched metadata description will support users, whether a DCO community member or not, to discover and access datasets of interest. By registering datasets from external repositories in the DCO data portal and enriching their annotation, the DCO data portal facilitates the reuse and circulation of existing data resources in the field of geosciences. Moreover, it leverages the existing data resources to create a unique data legacy for the global deep carbon-related research.

For researchers in the DCO community, by keeping their profile in DCvO up-to-date, adding publications, updating projects throughout the year for easier report generation, and adding datasets into the system, the researcher enables linking

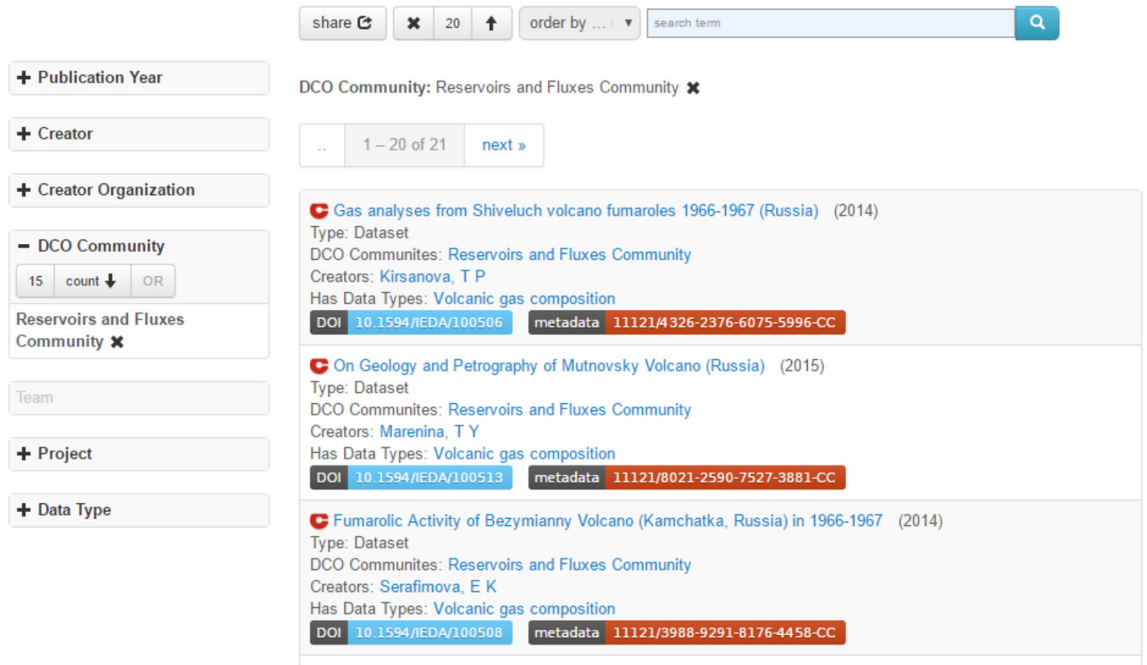

FIGURE 7 | DCO dataset records imported from external data repositories.

people, organizations, publications, events, projects, grants, data, and more. Whenever a researcher enters information into DCvO they are contributing an ever-increasing network of linked information and documentation of the legacies of the DCO.

DISCUSSION

With the introduction of the Web in the early 1990's the world realized the value of linking documents together and scientific research communities were among the first adopters. In Web 2.0 we saw a quick increase in the use of the Web for social media, e-commerce, and the dynamic creation of pages. With the advent of Semantic Web technologies (Berners-Lee et al., 2001), Web 3.0 became possible, a combination of principles and technology that enables us to link various objects and resources together in ways that both humans and computers can understand. In the Semantic Web, structure and interoperability of datasets are facilitated by the use and reuse of ontologies (Ma and Fox, 2013; Ma et al., 2014a). The reuse of ontologies on the one hand lower the technical burden for ontology maintenance, and on the other hand also promote the sustainability of the knowledge network in the DCvO.

The geoscience community is starting to understand that data, knowledge, and services are linked together, rather than distinct areas of focus or worse isolated. Such linkage is what the DCvO offers in providing dissemination of all outputs of the DCO both for the DCO science community, the broader geoscience community, and the general public. With a knowledge network of interconnected objects and resources we are able to facilitate more and better collaborations, find colleagues working on similar projects, contribute data that can be useful to others, discover methods, or tools that can analyze data in new ways, and study the patterns in scientific activities and outputs. The accessibility of resources in the DCO knowledge network is significantly improved by using the DCO-ID, which assigns a persistent and stable Web identifier to every resource that is part of DCvO. The DCO-ID will always be valid, even if the underlying resource changes. A researcher's profile is available via a DCO-ID and can be linked to external identifiers such as ORCID, ResearcherID, etc., Each publication of the researcher is available via a DCO-ID, in addition to a DOI if the DOI exists. Moreover, the datasets published by the researcher are also available via their DCO-ID. Our aim is to keep the resources in DCvO as a legacy of the DCO community, which will serve future works of deep carbon science even after the end of the DCO program.

The framework of DCvO presented in this paper, to our knowledge, is the first attempt to connect the platforms Drupal, VIVO, CKAN, and the Handle system. The aim is to leverage the advantage of each platform, such as Drupal for the front end, VIVO for object registration and connection, CKAN for data deposit and the Handle system for persistent identifiers. Many works in this framework was the first technical practice. Data and service standards developed by communities such as those W3C ontologies significantly reduced burdens in the technological development and implementation. Other standards, such as those developed by the Open Geospatial Consortium and the International Organization for Standardization (ISO), although not being described in detail in this paper, also play essential roles in the data flow within the DCvO framework and the communication between DCvO and external systems. The built

system is modularized and is highly reusable. The framework, or parts of it, can be easily deployed at a program or institutional level to address needs similar to DCvO.

There are already data resources on the Web that provide structured information for focused topics such as publications, datasets, geologic samples, and researchers. Some of those resources provide information service via persistent identifiers such as DOI. Such information can be a significant contribution to the DCO knowledge network. The organization CrossRef (http://www.crossref.org) provides an interface where machines can retrieve the metadata record of publications using their DOI. DataCite (https://www.datacite.org) has been providing a similar service for datasets (sometimes called data publications). ORCID (https://orcid.org) has been working on structured information for uniquely identifying researchers, and the International Geo Sample Number (IGSN)'s focus is on the structured records and identification of physical samples. One can see that in the future it will be easier to access detailed and authentic information of an object using a persistent identifier. Such capability will unlock individual records that can become part of many knowledge networks. Nevertheless, there is still the challenge to weave those various topics of information into a specific network such as the DCO knowledge network. Most data resources on the Web, such as those mentioned above, are only able to provide records in literal values, though those values may be structured and well-organized. In contrast, in the knowledge network each object or resource is regarded as a node in the network. To match literal values to nodes in the knowledge network we have developed some semi-automatic techniques in DCvO. For example, one function is matching an author from a CrossRef record to a person in the DCvO. Due to the issues of same names or using initial for first name, a data curator will verify the matching results and choose the correct match. A remaining challenge is to robustly recognize entities from literal values, address ever-present ambiguities in records and weave them into a knowledge network. There is still work to be done.

CONCLUSION

The work presented in this paper gives a glimpse at one of the first initiatives that deployed leading-edge Semantic Web technologies for large, international research collaborations in the geoscience community. The aim for DCvO is to create more than just a data repository, but a knowledge portal. By using a knowledge network underpinned by ontologies and leveraging state-of-the-art methods in data stewardship, DCvO is able to support various aspects of collaborative geoscience research. The information in DCvO were collected from both the DCO community contributions and several extramural data resources. Detailed records were stored in a way that both humans and machines can read and understand. In the environment of the Semantic Web, a key feature of DCvO is the inter-connections among various registered objects and resources as well as the flexible ways to discover and access them. With the knowledge network the DCO community members are able to add publications and datasets that can be useful to others, find colleagues working on similar projects, discover methods and tools that can be used to analyze data in new ways, and create more and better research collaborations.

AUTHOR CONTRIBUTIONS

XM participated in the research and led the writing of the manuscript. PW, SZ, JE, AE, YC, HW, and HZ. participated in the research and contributed to the manuscript writing. PF. led the research and contributed to the manuscript writing.

ACKNOWLEDGMENTS

This work was funded by Alfred P. Sloan Foundation through the Deep Carbon Observatory [Award numbers: APS: 2012-10-02 (RPI) and APS: 2014-06-02 (RPI)].

REFERENCES

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web. *Sci. Am.* 284, 34–43. doi: 10.1038/scientificamerican0501-34

Devaraju, A., Klump, J., Cox, S. J., and Golodoniuc, P. (2016). Representing and publishing physical sample descriptions. *Comput. Geosci.* 96, 1–10. doi: 10.1016/j.cageo.2016.07.018

Ellis, D., and Vasconcelos, A. (1999). Ranganathan and the net: using facet analysis to search and organise the World Wide Web. *Aslib Proc.* 51, 3–10. doi: 10.1108/EUM0000000006956

Fox, P. (2015). "Why we need to get smart about data to be better stewards: making smarter virtual observatories," in *2015 IEEE International Geoscience and Remote Sensing Symposium* (Milian: IGARSS), 1351–1353.

Fox, P., and McGuinness, D. L. (2008). *TWC Semantic Web Technology*. http://tw.rpi.edu/web/doc/TWC_SemanticWebMethodology (Accessed 05.04.17).

Glaves, H. (2017). "Developing a common global framework for marine data management," in *Oceanographic and Marine Cross-Domain Data Management for Sustainable Development,* eds P. Diviacco, A. Leadbetter, and H. Glaves (Hershey, PA: IGI Global), 47–68.

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum Comput. Stud.* 43, 907–928. doi: 10.1006/ijhc.1995.1081

Harnad, S., and Brody, T. (2004). Comparing the impact of open access (OA) vs. non-OA articles in the same journals. *D-lib Mag.* 10, doi: 10.1045/june2004-harnad

Hey, T., and Payne, M. C. (2015). Open science decoded. *Nat. Phys.* 11, 367–369. doi: 10.1038/nphys3313

Lebo, T., Sahoo, S., and McGuinness, D. (2013). *PROV-O: The PROV Ontology*. Accessible online at: http://www.w3.org/TR/prov-o/

Lehnert, K. A., Vinayagamoorthy, S., Djapic, B., Klump, J. (2006). "The digital sample: metadata, unique identification, and links to data and publications," in *American Geophysical Union Fall Meeting 2006* (San Francisco, CA) Abstract # IN53C-07.

Ma, X., Erickson, J. S., Zednik, S., West, P., and Fox, P. (2016). Semantic specification of data types for a world of open data. *ISPRS Int. J. Geo Inf.* 5:38. doi: 10.3390/ijgi5030038

Ma, X., and Fox, P. (2013). Recent progress on geologic time ontologies and considerations for future works. *Earth Sci. Inform.* 6, 31–46. doi: 10.1007/s12145-013-0110-x

Ma, X., Fox, P., Rozell, E., West, P., and Zednik, S. (2014a). Ontology dynamics in a data life cycle: challenges and recommendations from a Geoscience Perspective. *J. Earth Sci.* 25, 407–412. doi: 10.1007/s12583-014-0408-8

Ma, X., Fox, P., Tilmes, C., Jacobs, K., and Waple, A. (2014b). Capturing provenance of global change information. *Nat. Clim. Change* 4, 409–413. doi: 10.1038/nclimate2141

Ma, X., West, P., Erickson, J., Zednik, S., Chen, Y., Wang, H., et al. (2015). "From data portal to knowledge portal: Leveraging semantic technologies to support interdisciplinary studies," in *Proceedings of the Diversity++ Workshop at ISWC 2015*, (Bethlehem), 6

Ma, X., Zheng, J. G., Goldstein, J., Zednik, S., Fu, L., Duggan, B., et al. (2014c). Ontology engineering in provenance enablement for the National Climate Assessment. *Environ. Model. Softw.* 61, 191–205. doi: 10.1016/j.envsoft.2014.08.002

Maali, F., and Erickson, J. (2014). *Data Catalog Vocabulary (DCAT)*. Available online at: http://www.w3.org/TR/vocab-dcat/

Mitchell, S., Chen, S., Ahmed, M., Lowe, B., Markes, P., Rejack, N., et al. (2011). "The VIVO ontology: enabling networking of scientists," in *Proceedings of the ACM WebSci'11* (Koblenz), 2.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al. (2015). Promoting an open research culture. *Science* 348, 1422–1425. doi: 10.1126/science.aab2374

Seo, C., Lee, S. W., and Kim, H. J. (2003). An efficient inverted index technique for XML documents using RDBMS. *Inf. Softw. Technol.* 45, 11–22. doi: 10.1016/S0950-5849(02)00157-X

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer AS and handling Editor declared their shared affiliation, and the handling Editor states that the process nevertheless met the standards of a fair and objective review.