RDA ADOPTION PROJECT REPORT

Deploying Research Data Alliance Data Type Registry and Persistent Identifier Information Types in the Deep Carbon Observatory Data Portal

Xiaogang Ma, John S. Erickson, Stephan Zednik, Patrick West, Peter Fox
Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th Street, Troy NY 12180, USA
Email: Xiaogang Ma (max7@rpi.edu); John Erickson (erickj4@rpi.edu); Stephan Zednik (zednis2@rpi.edu) Patrick West (westp@rpi.edu); Peter Fox (pfox@cs.rpi.edu)

Executive summary

The Research Data Alliance (RDA) - Data Type Registry (DTR) Working Group addresses a part of a core problem relevant to interoperability among data management systems: the ability to parse, understand, and potentially reuse data retrieved from others. The RDA - Persistent Identifier Information Types (PIT) Working Group addresses the essential types of information associated with persistent identifiers. During the period from February to August 2015, we have undertaken an effort to adopt the DTR and PIT outputs in the Data Portal of the Deep Carbon Observatory (DCO) and have received positive results.

THe RDA DTR and PIT outputs (hereafter referred to in brief as DTR and PIT) serve various data consumers, including both humans and machines, in a compelling way: given a dataset identifier, discover detailed information about the structures and meanings of that dataset, and act accordingly. The data type issue has traditionally been addressed at the syntactic level, e.g., file types, MIME types, and so forth, but not at the semantic level, e.g., the data in a table column are all integers but what do they represent? DTR explored ways to enable data creators to record and make known the implicit assumptions of a dataset. PIT developed a conceptual model for structuring typed information and deployed an application programming interface for access to typed information. The PIT deliverables were tested using a demonstrator system that is shared with the DTR deliverables.

DCO has implemented a centrally-managed digital object identification, object registration and metadata management service known as the DCO Data Portal. The Portal provides the digital object registration process for DCO Community members, composed of two key elements: 1) DCO-ID handle generation based on the global Handle System and 2) metadata collection for each registered object. The DCO Data Portal is maintained by the DCO Data Science Team at the Tetherless World Constellation of Rensselaer Polytechnic Institute. Datasets generated by the DCO Community cover various formats and topics in Earth and space sciences, and we have encountered a large number of digital object registrations. The curation and reuse of registered datasets within the DCO Data Portal was well suited for testing deployments of RDA DTR and PIT.

RDA DTR and PIT include both methodology and technical solutions in the demonstration systems. Although the demonstration systems require further adaptation and evolution to ensure they are ready for production, the methodologies embodied by DTR and PIT are highly recommended for data repositories of various disciplines. Stakeholders should be encouraged to incorporate implementations of DTR and PIT into current infrastructures. Such efforts will significantly facilitate work on data curation and promote the sharing and usability of deposited data.

1 Background

The background of this adoption research is the reusability of data. We can describe such a case by using a scenario of researchers using resources from the open data environment on the Web. If the researchers retrieve a file from the Web and want to use it, they will need to know the format, structure, parameters and meaning of the data, and perhaps also the tools and services that can be used to process the data. In the world of open data, the researchers receive no direct support or help from the data producers, which indicates that the metadata of the retrieved data is often the only source for the information that the researchers need.

The ‘data types’ registered in a DTR are meant for explicating assumptions or information inherent in data. Once registered, ‘data types’ will be assigned unique identifiers (handles). A registered data type is assumed to be resolvable to some useful (by human or machine) "explanation" and "definition" of that type. There will be multiple DTR instances, and each governed by its own project, group or community. All those DTR instances reuse some common basic types, which are called ‘primitives’. Those primitives (e.g., basic data types) will be registered in a type registry presumably managed by the Corporation for National Research Initiatives (CNRI). So, we can expect a two level hierarchical federation of the DTR. The higher level is a list of primitives and the lower level is the specific data types defined within a DTR. PIT was formulated under the goal of harmonizing the basic information types associated with persistent identifiers (PIDs). It currently uses a property-type-profile model. In this model, every PID consists of a number of properties. Every property also bears a PID and its essential elements are a name, a range and a value. Every property is registered in a DTR. The registered information includes property range, name and additional provenance information. Every type is registered in the DTR and bears a PID. A type consists of a number of properties and additional descriptions, and provenance information. A profile consists of several types. A profile also bears a PID if it provides mandatory properties of all types in the profile.

The primitives in a DTR are comparable to a list of defined data type classes in the DCO ontology, such as Dataset, Image, Video, and Audio, etc. The properties associated with each PID information type in PIT are comparable to the properties associated with those data type classes in the DCO ontology. Currently, a registered DCO dataset is regarded as an instance of one of those classes. We can see the potential to further annotate a registered dataset with the specific data types defined within a DTR, and each data type has a PID (i.e., mechanism in current DTR and PIT demos). For example, we add a "dco:hasDataType" property in the DCO ontology, and use it to assert DTR data types for a registered dataset. The specific data types of a dataset may be multi-valued, that is, a given data object could have multiple assertions of dco:hasDataType, which could be from the DCO data type registry or from some external data type registry that the maintainer knows about. Such annotation will increase the

understandability of data objects at the semantic level. For instance, an external service querying the "type" of a DCO-ID identified object would receive a list of values associated with dco:hasDataType.

The DCO endorsed data policies for establishing the framework for the long-term stewardship of carbon data and information. To make the "network" aspect of DCO work, contributors to and users of the DCO need to comply with a set of guidelines or "norms of behavior." The norms assure users of the DCO that contributors have provided, to the best of their ability, high-quality data, information, and other digital resources about carbon in Earth's interior, accurately described according to agreed standards, such as classes and properties in the DCO ontology.

2 Methods, Implementation and Results

2.1 Data types

The differentiation between the basic data type and specific data type, as a method derived from the RDA DTR deliverables, paves the way for data type description in the DCO data portal. The environment of this work is the Semantic Web, which is defined as an extension to the original World Wide Web such that meanings of data are made machine readable. Among the various ontologies available in the Semantic Web, there are already defined classes for types of data resources. For example, the DCMI Type Vocabulary defines a list of types, including Collection, Dataset, Event, Image, Interactive Resource, Moving Image, Physical Object, Service, Software, Sound, Still Image, and Text. Each DCMI Type is defined as an owl:Class. A resource in our data repository can be asserted as an instance of one of those classes. For instance, <dco:data_001 rdf:type dctype:Dataset>. In other ontologies there are also similar classes for resource types such as vivo:Dataset and vivo:Video in the VIVO Ontology and bibo:Code in the Bibliographic Ontology.

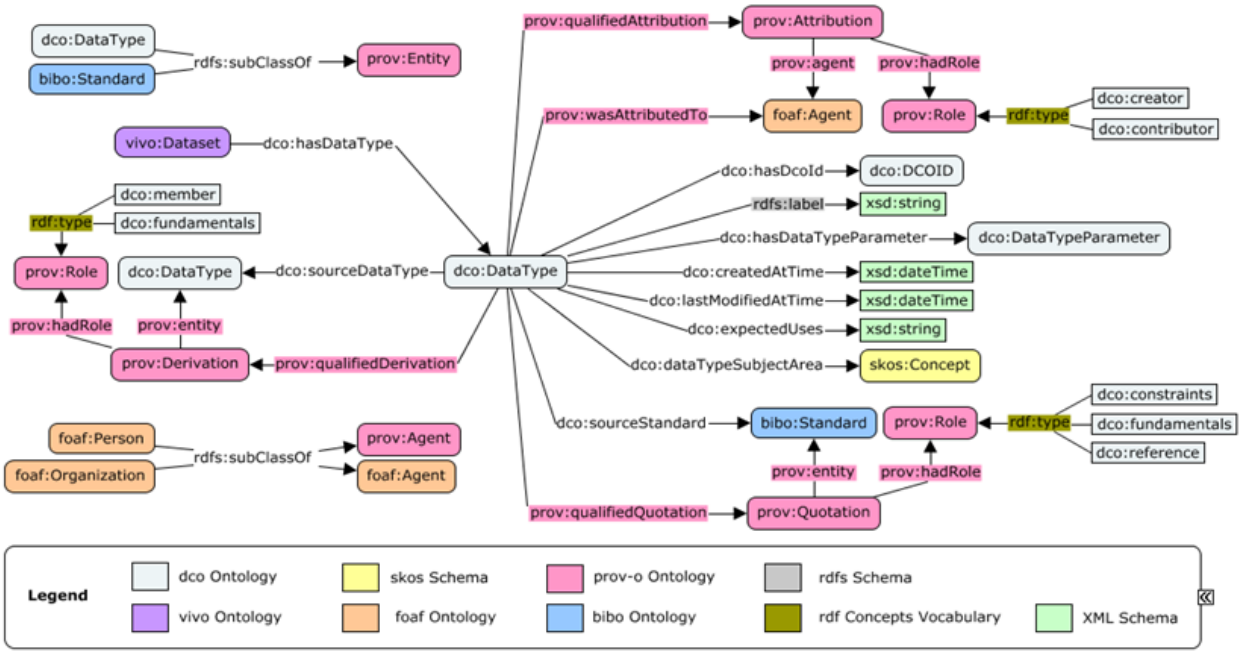

Figure 1 A conceptual model for the specification of data types.

In the sense of ontology modeling, we regard data as a general concept that cover not just instances of dctype:Dataset but also instances of other classes, including those defined in the DCMI Type Vocabulary and other ontologies. We call those resource type classes "basic data types" because each of them categorizes the nature of a resource. For a data resource instance, we can know its basic data type by reading the triple with predicate rdf:type, such as the above example <dco:data_001 rdf:type dctype:Dataset>. Nevertheless, the basic data type offers only limited information for understanding the meaning of resource. That is why we took up the work on the formal specification of data types, and we call them "specific data types". The specific data type provide more information about a data resource. For example, we can use the triple <dco:data_001 dco:hasDataType dco:volcanicGas> to annotate a specific data type. Here the dco:volcanicGas is not a keyword or tag. Instead, it is an instance of the specific data type class dco:DataType, and there are a group of properties describing it. Figure 1 shows the datatype properties and object properties associated with the class dco:DataType in our designed model for data type specification. The model demonstrates that specific data types can be used to annotate instances of vivo:Dataset, and in practice it can also be used to annotate other data resources such as instances of dctype:Image, dctype:Sound, etc. The model was made a part of the ontology we developed for the data portal of the DCO community, and used a few properties and classes from that ontology.

2.2 Persistent identifiers

The DCO-ID shares ideas in common with the RDA PIT. The DCO-ID uses the Global Handle System (GHS) to assign a persistent unique identifier to almost every object in the DCO Data Portal. The DCO-ID is similar to the Digital Object Identifier (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 (Figure 2). In this way we can keep a persistent and stable legacy for the activities and outputs of the DCO community.

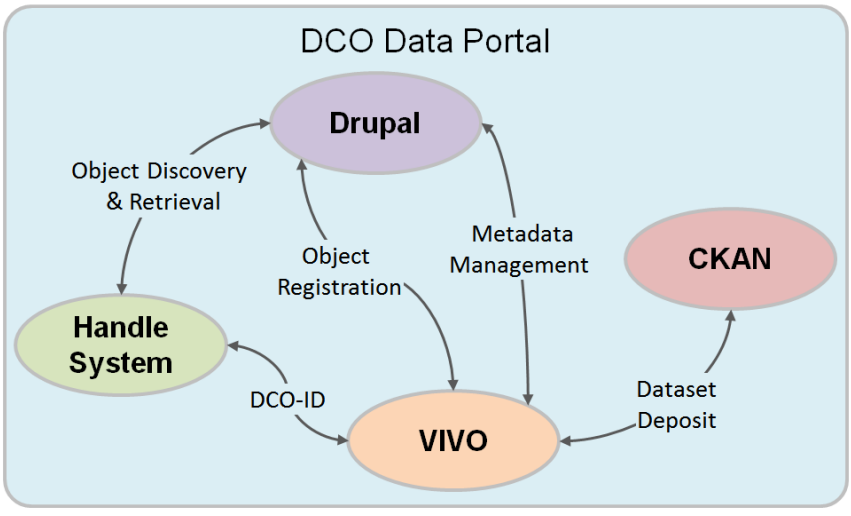

Figure 2 DCO-ID as a mechanism of persistent identifier for both object registration and retrieval.

Each DCO-ID can redirect to the Web profile (often a landing page) of an object, where the detailed metadata can be found. In the DCO Data Portal each object is the instance of a class. The metadata items describing an instance are properties. All those classes and properties are organized by the DCO ontology. The classes and properties behind the instance and DCO-ID are comparable to the information types proposed in the RDA PIT. The DCO-ID allows both humans and machines to retrieve metadata of any object deposited in the data portal.

2.3 Results of implementation

With the conceptual model of data type and the persistent identification enabled by the DCO-ID, we developed functions of data type registration, identifier allocation, and data type browsing in the DCO data portal. For the data type registration we used the default user interface of the VIVO platform, which follows the general workflow of creating an instance for any class. Once a data type instance is created, the data portal assigns a unique DCO-ID for it. Then on the VIVO profile of the data type, a user fills in records for the properties describing the data type. Figure 3 shows a part of the profile of the registered data type "Thermodynamics of chemicals and minerals".

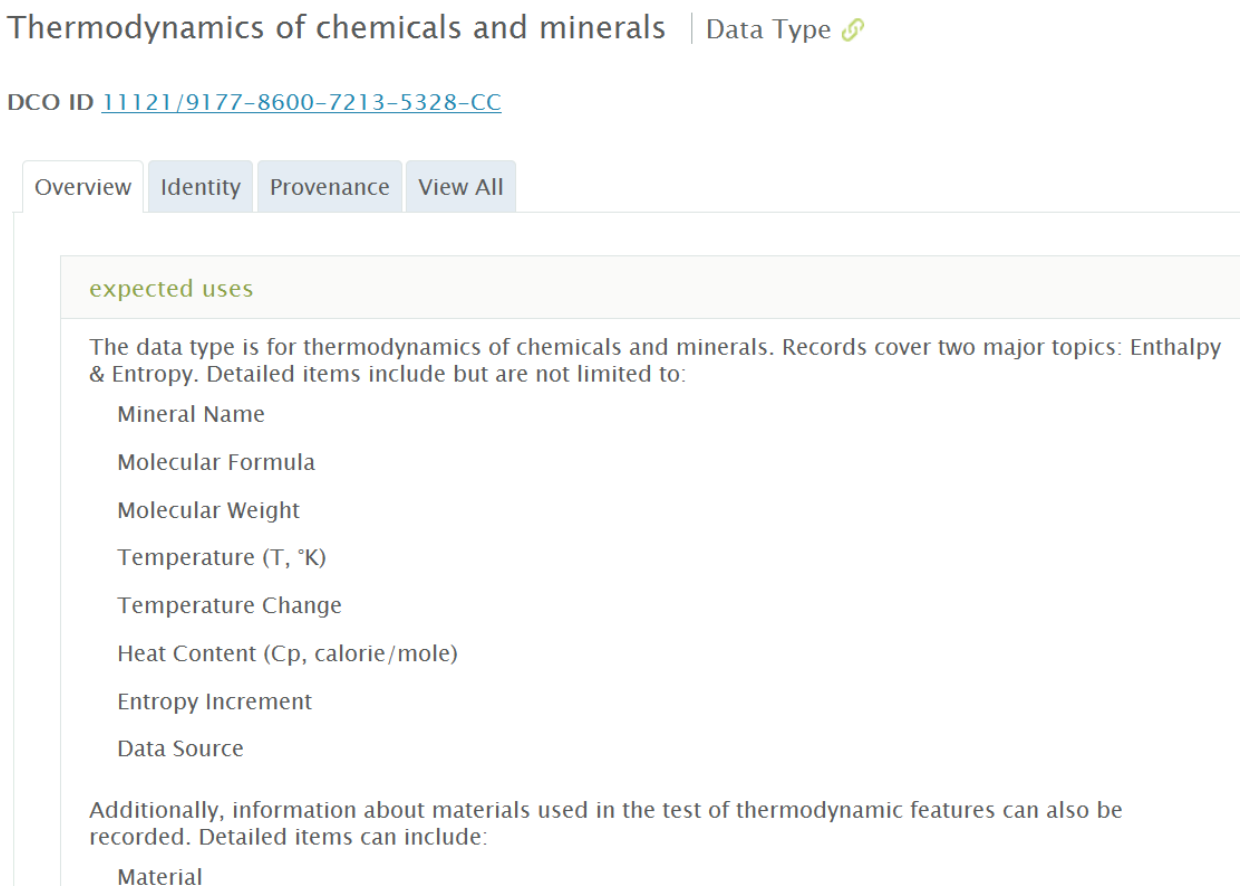

Figure 3 Screenshot of the profile of a registered data type.

We also developed a faceted browser for all the registered data types. Figure 4 shows a screenshot of the faceted browser with a few data types we registered as test examples. On the left of the user interface

there is a list of facets, which are related to the corresponding properties of registered data types in the portal. A user can search among the data types by choosing records in those facets. By clicking the DCO-ID of a data type the user can go to the profile of that data type for detailed metadata. A feature of the browser is that, once the chosen records in a facet are changed, all records in other facets as well as the data type results will change correspondingly. The user can make selections in several facets to search for one or more certain data types.

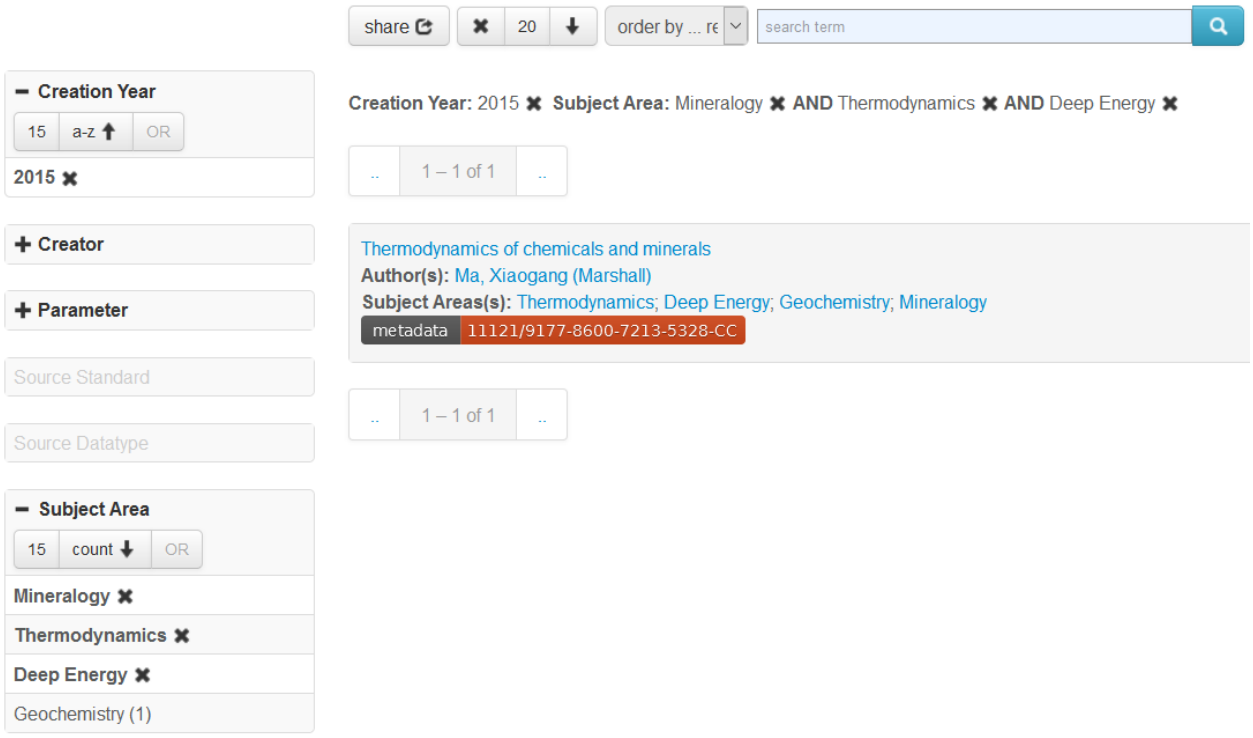

Figure 4 A faceted browser for registered data types.

3. Evaluation and Recommendations

Science today is increasingly facilitated by Open Data. Our adoption of the RDA DTR and PIT outputs enabled us to review and update "type" functionalities in the DCO Data Portal. The work on the conceptual model of data type enriches the semantic description of datasets, and the work on DCO-ID allows a persistent and unique identifier for various objects deposited in the DCO Data Portal. The information in such description is not only human readable but also machine readable, which will provide a valuable resource to people who access and use datasets that they did not generate (data science)..

In summary, we adopted the methodologies proposed by the DTR and PIT working groups and made adaptations on the technical approaches to make the results an integrated part of the DCO Data Portal. Our differentiation of basic data types and specific data types are derived from the primitives and specific data types developed by the RDA DTR working group. In the implementation, we utilized semantic technologies. From the point of view of ontology engineering, both primitives and specific data types in the DTR design are at the instance level, i.e., they both are registered data types. In our work, the basic data types are at the class level, i.e., they are classes in an ontology, and data resources are instances of

them. The specific data types in our work are at the instance level, i.e., they are all instances of the class "dco:DataType". If we put the specific data types at the class level, then we need to update the ontology frequently to include the new data type classes created. Moreover, the way to use a specific data type class is to make data resource an instance of it. Using a single class "dco:DataType" and making all specific data types as instances of it significantly reduces the efforts needed to update the ontology and to maintain the framework of the data portal.

The DCO-ID implements the GHS to make the registered identifiers persistent. The information types proposed in the RDA PIT working group were realized by updates to the DCO ontology. That is, each DCO-ID represents a DCO object, which is the instance of a class in the ontology and is described by a number of properties. Because the DCO Data Portal is underpinned by ontologies, after updating classes and properties in the DCO ontology, we deployed them in the DCO Data Portal quickly and smoothly. This shows the advantage of semantic technologies. In our adoption work the ontology is comparable to the centralized portal for identifier information types proposed by the RDA PIT working group. By reusing existing community ontologies such as those for publications, datasets, organizations, etc. we improved the interoperability of metadata registered in the data portal. Therefore, though our technical approach for persistent identifiers and their information types is slightly different from the PIT working group, we share the same objective and our work benefited from the methodology of PIT, that is, the machine accessibility and readability of information.

The NSF RDA/US funding for this project supported graduate research assistants, Apurva Kumar Sinha and Congrui Li as well as staff time and travel for RDA events, and enabled us to complete the proposed work within the timeline (February 01 - August 31, 2015). Being members in the RDA DTR and PIT working groups and having access to the group documents on the RDA website is a great advantage that allows us to know more details about the background, methodology as well as technology for the implementation of the deliverables of the two groups. The two RDA plenary meetings in 2015 were opportune times for us to present the initial and final results, respectively. Moreover, the environment of the DCO data portal set up the foundation for both flexible and quick adoption. With a view toward the future, meaningful data types and persistent identifiers is receiving increasing acceptance in the data science community. Our work demonstrates that the methodologies of both RDA DTR and PIT are highly implementable, especially in our platform environment based on the Semantic Web. Also, the technical framework in the current demonstration systems of DTR and PIT can be adapted or further extended for production uses. We look forward to consulting with other interested adopters of DTR and PIT to share experiences, details, ontologies, and documentation/ codes. Please visit http://tw.rpi.edu/web/project/RDAAdoption for details on this and future work.

A few publications resulted from this project:

1. Ma, X., Erickson, J.S., Zednik, S., West, P., Fox, P., Formal semantic specification of data type for a world of open data. Journal paper manuscript to be submitted.
2. Ma, X., West, P., Erickson, J., Zednik, S., Chen, Y., Wang H., Zhong H., Fox, P., 2015. From data portal to knowledge portal: Leveraging semantic technologies to support interdisciplinary studies. In: Proceedings of the Diversity++ Workshop at ISWC 2015, Bethlehem, PA. 6 pp. http://tw.rpi.edu/web/doc/2015_Diversity_DCO.pptx
3. Ma, X., Erickson, J.S., West, P., Zednik, S., Fox, P., DCO-DS team, 2015. Adoption of RDA DTR and PID in the Deep Carbon Observatory Data Portal. RDA Sixth Plenary Meeting, Paris, France. Oral Presentation by P. Fox. http://tw.rpi.edu/web/doc/20150921_slides_RDA_P6.pptx

4. Ma, X., Erickson, J.S., West, P., Zednik, S., Fox, P., Wang, H., Chen, Y. 2015. Formal Specification of Data Types in the Deep Carbon Observatory Data Portal. DCO International Science Meeting 2015, Munich, Germany. Poster. http://tw.rpi.edu/web/doc/xmadco2015
5. Ma, X., Erickson, J., West, P., Zednik, S., Fox, P., 2015. Data types and persistent identifiers in the Deep Carbon Observatory Data Portal. RDA Fifth Plenary Meeting, San Diego, CA. Poster. http://tw.rpi.edu/web/doc/rda-p5-xm
6. Zednik, S., Ma, X., Erickson, J., West, P., Fox, P., 2015. Adoption of RDA DTR and PID in Deep Carbon Observatory Data Portal. RDA Fifth Plenary Meeting, San Diego, CA. Oral Presentation. http://tw.rpi.edu/web/doc/rdap5-adoption-dco