Data Type Registries (Phase 2) Working Group Case Statement

Background

Interpretation and use of scientific datasets by those that are not engaged in the creation or production of those datasets is pivotal for enabling science that is driven by data. RDA has made significant progress in addressing this issue through the initial Data Type Registries (DTR) WG, which has now finished, but much remains to be done, hence this proposal for a follow-on working group. The fundamental effort here is to describe scientific datasets in a human-and-machine-readable fashion, enabling humans and software to parse and understand the semantics, context, and assumptions behind the data. We reference all such descriptions "data type records", regardless of the standard or best practices standing behind those descriptions. Data types complement traditional descriptive metadata records, providing re-usable descriptions of dataset structure and semantics aimed mainly at supporting data processing, while at the same time providing an additional attribute that can be used for a certain kind of discovery. The initial DTR WG focused on developing an infrastructural component that would manage such data type records. The new WG, DTR2, will focus on aiding data producers come up with useful data type records.

Executive Summary

Multiple standards that exist today, such as RDF/OWL and ISO 11179, apply equally to this space of creating data type records to aid data consumers. Standard schema languages are available for some data formats (XML, JSON, proposed for CSV). There are also some more specific precedents, such as Earth Science Markup Language [1] and Ecological Metadata Language [2] that provide methods to describe the internal structure of datasets exchanged in certain technical communities. While each of those standards and best practices benefit specific sets of communities, we believe that there are common themes and analyses that data producers follow in general to produce data type records regardless of the standard they choose for representing this information. Open Knowledge, for example, is addressing the issue through their Data Packages [3] proposal, but the data type descriptions are currently not re-usable or registered externally. The goal of the DTR2 WG is to identify those common themes across communities and to subsequently produce a set of guidelines for communities to identify ***what*** information should be provided to enable third parties within and potentially also outside their communities to easily use their datasets. We call such guidelines "recipes". A core piece of the proposed solution is the assignment of a unique and ideally persistent identifier to a data type record. Such an identifier, aka a Data Type ID, may then act as a reliable link or

[1] http://projects.itsc.uah.edu/esml/
[2] https://knb.ecoinformatics.org/#external//emlparser/docs/index.html
[3] http://data.okfn.org/doc/data-package

shortcut to associate datasets with their respective types. Once a recipe is followed, a given community can choose to represent the identified information (i.e., data type record) in any one or multiple existing standards[4]. We do not constrain 'community' here; in this context a community is defined by its intention to share one or more data type recipes, so there will often be multiple recipes in scope for an application. But at least if they are registered and available the process of convergence is enabled. The registration paradigm is an enabler, not an enforcer.

The DTR2 WG would function for 18 months starting upon WG approval. The WG will extend the initial Data Type Registries (DTR) WG operated within RDA by making use of the Data Type Registry for its ongoing activities. During the operational period, the WG will engage with various data producers and consumers to identify recipes for as many kinds of datasets (tabular, hierarchical, image, etc.) as possible and to explicate the different varieties of information (structural information, service information, vocabulary information, etc.) that dataset consumers would find useful. The number of dataset kinds and information varieties that the WG will produce recipes for will be constrained only by time and member interest.

The WG will also focus on creating relationships across data type records in order to leverage previously created data type records. To that end, the Data Type Registry software produced by CNRI as part of its DTR WG activity will be used to demonstrate the relationship functionality across data type records. Furthermore, federation across multiple instances of Data Type Registries, including alternate registry implementations, will be examined. Such federation will aid in the reuse of data type records across communities. Specifically, policies and governing models found to be useful for federating data type records across organizations boundaries will be identified and reported.

The WG will produce a prototype that implements some or all of the recipes produced. The WG will aid some of the non co-chair institutions to experiment with and adopt the produced output, e.g., NIST laboratories, Deep Carbon Observatory project, International DOI Foundation, etc., all of whom are potential adopters of the proposed WG output.

Work Plan

Based on our previous experience with the Data Type Registries WG, it is clear that community requirements vary in regards to the details of data type records that will be useful to their likely consumers. A given community may wish to focus only on structural aspects of data; another community might wish to express semantics behind its data rather than the structural aspects; a third community might choose

[4] It is our hope at this point that communities will converge to using a common representation for specifying core information about any dataset. This WG would specify a representation for describing "core" information, but would not prescribe such representation.

to express both; a fourth may wish to express constraints behind the data; and so on. Therefore, a single-purpose data characterization approach will not work even if there are only a handful of communities involved. To this end, this WG will identify recipes for describing data for a limited number of use cases. By recipes, we mean a set of guidelines, e.g., specify values for these five elements and here are what usually is captured in those five elements, and so on. If those recipes are followed, a given community will have defined data types that should make consuming data easier, especially by consumers not involved in the data creation process.

The WG will extend the work performed by the Data Type Registries working group to enable "Data Type ID" as a linking mechanism or shortcut to one or more data characterizations, aka data types. In addition to producing recipes for identifying information for characterizing different aspects of data, the WG will also identify a core set of elements that will be used to describe the generic aspects of any data set. Furthermore, and if found to be technically and practically feasible, the working group will define a way to build more complex data types from existing simpler types. To this end, the working group will try to define primitive types that may become the foundation for defining complex types by any community.

Figure 1 illustrates the scope, the primary set of communities, and the use cases that will be considered for this exercise.

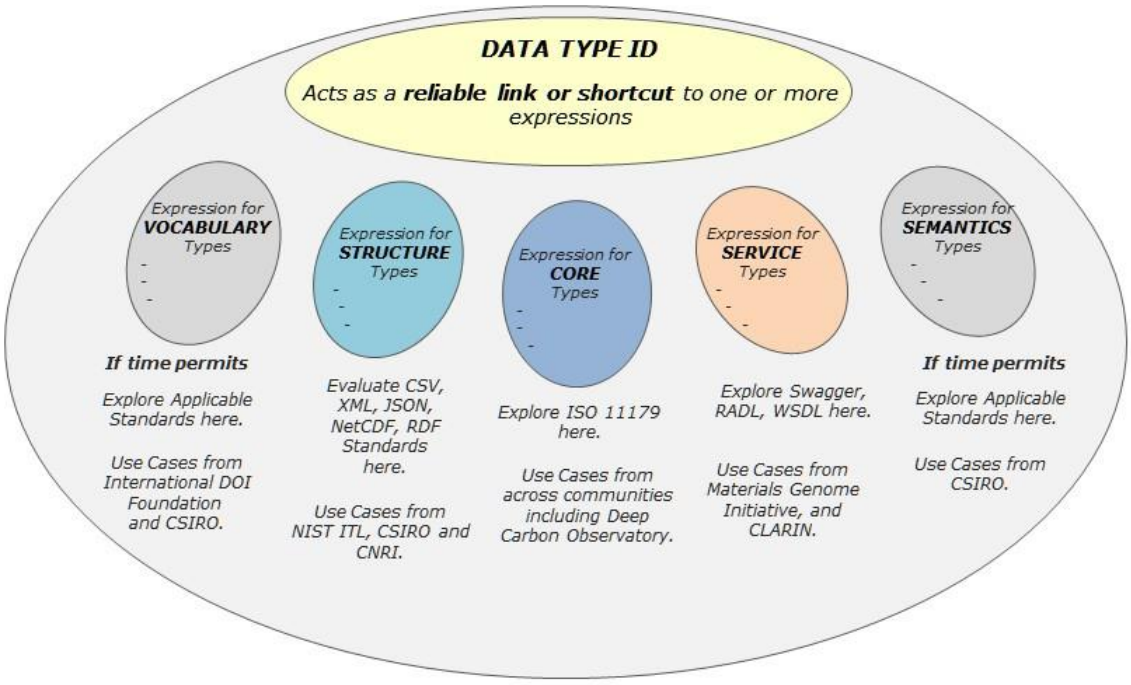

Figure 1: Use Cases and Communities Considered in the Working Group

The use cases that will be considered in this working group will be from groups that would like to characterize structural information (schemas, columns, etc.), record definitions/vocabulary, and link to (Internet/web) services. Performing this work in

the context of RDA allows us to access requirements from a range of disciplines and communities, and thus develop methods and solutions that may be quite generally applicable.

The WG will produce recipes for typing data – at least one recipe per use case plus one recipe for producing a core description. It is important to note that a given community may choose to combine two or more recipes to characterize a given dataset.

The WG will consider and include standard approaches and specifications to the extent possible for serializing the type record as a whole and also for expressing and serializing each of the type variations.

The WG will consider and define approaches for specifying relationships between data types in order to enable reuse of previously defined data types.

The WG will also work with participating communities to identify and report at least one governance strategy for federating data types across various instances of the Data Type Registries (DTR) run by different communities. Questions related to whether a data type can be modified or altered, if so by whom, and which data types are propagated to other DTR, if so under what circumstances, and who moderates those data types, and for how long, etc., will be discussed.

Finally we will work with and encourage cross-fertilization with other RDA WGs and IGs. DTR2 is intended to be useful as low level infrastructure and as such is potentially of interest to most technically-oriented WGs and IGs and most WGs and IGs are potential sources of use cases for Data Type Registries. That said, there are few groups to which we will try to pay special attention and with which we hope to work. The PID Information Types WG was tightly connected with DTR and utilized DTR for registering types. It is now closed but the core model is now being used by the Research Data Collections WG and that group is potential DTR user and its use cases will provide a useful perspective. A number of domain groups have also identified DTR as a potential resource for their work. These include the Empirical Humanities Metadata WG, the International Materials Resource Registries WG, and the BioSharing Registry WG. Joint sessions with these groups have been held at past or are planned for future plenaries. There is also a clear connection with the Metadata Standards Catalog WG as data type is a type of metadata and may become part of such a catalog. DTR also has a significant connection with Vocabulary Services, as data-type descriptions will use vocabularies, and specialized vocabularies might use data-types. Finally we note that as one of the initial core infrastructure groups the DTR is part of the Data Fabric Interest Group initial configuration effort and we anticipate useful interaction with ongoing DFIG efforts at pulling together various RDA outputs.

Adoption Plan

CNRI has been working with National Institute of Standards and Technology (NIST) laboratories and other US federal agencies for the last two years to aid those

agencies describe datasets. NIST, for example, has been evaluating the Data Type Registry software that CNRI produced as part of its DTR WG activity. Some of the technical questions and challenges that surfaced in the NIST evaluation would be addressed in this WG. It is highly likely that NIST will adopt the recommendations and outputs from this WG to further their experimentation and adoption efforts. It is also likely that other federal agencies will evaluate and adopt the outputs from this WG.

CNRI has also been engaged with EUDAT. In particular, the DTR software is now being evaluated by EUDAT for use in its research and production activities. EUDAT is currently consolidating its identifier services and a DTR instance may become an additional service building block targeted at the scientific EUDAT user communities. It is likely that the outputs from this WG will be evaluated and adopted, as these outputs will add incremental value to the outputs produced by the DTR WG.

DKRZ is currently developing an identifier and object management approach for the upcoming scientific experiment data of the Coupled Model Intercomparison Project (CMIP) coordinated by the World Climate Research Programme's Group on Coupled Modeling (WCRP-WGCM). One part of these activities is to describe and store the most essential information that characterizes the data objects managed by automated services. This can lead to the definition of one of the aforementioned community-specific recipes. CMIP data products are also subject to various processing workflows, for example through a suite of Web Processing Services co-developed at DKRZ. Since the data space is rather diverse and complex, introducing precise data typing and binding specific services to types can increase the automation and usability of these workflows. DKRZ will evaluate the DTR software and the WG outputs for usage within these services.

CSIRO Australia is the lead technical agency in the development of a national data architecture called 'OzNome'. A key aspect of this is to provide platforms to publish existing datasets for re-use. The ability to annotate technical datasets with data types will be an important enabling technology and will provide a useful platform for testing the DTR approach and software.

One outcome of the RDA working group PID information types (PIT-WG) was a preliminary classification of a first set of types intended for use as additional metadata stored inside the PID record. The ePIC Persistent Identifier Consortium for eResearch has as its main objective providing persistent identifiers for data management and will continue the aforementioned work of the PIT-WG together with an adoption of the outcomes produced by the DTR WG in the direction of federated DTRs. The focus will be to define a set of types that contain the information needed for automated data management services like data access, data curation, data versioning, data publication and similar processes with a high need of automation

Timeframe

18 months with a start date coinciding with WG approval.

Deliverables

The working group aims to produce the following deliverables:

1. Recipes for typing data for select use cases: types for defining structure, vocabulary, and types for linking to Internet services.
    a. Such recipes may include the notion of primitive data types and complex data types for purposes of reusing and evolving existing data types.
2. Governance policies for cross-community federation.

Select List of Institutions Participating Institutional Members

- NIST Information Technology Laboratory
- CLARIN Project Representatives
- Deep Carbon Observatory Representatives
- Materials Genome Initiative members
- Woods Hole Oceanographic Institute
- International DOI Foundation
- CSIRO Australia
- ePIC Persistent Identifier Consortium
- German Climate Computing Center (DKRZ)

Individual Members: We propose to simply keep the membership list from DTR.

Proposed Leadership: Simon Cox, Tobias Weigel, and Larry Lannom. This was both posted to the DTR list and explained at the Tokyo BoF. No objections were raised.