Data Type Registries

Proposed Case Statement for an RDA Working Group

Draft 0.5

4 January 2013

Larry Lannom, Daan Broeder

Background

Automated processing of large amounts of scientific data, especially across domains, requires that the data can be parsed without human intervention. Within a given domain that functionality can simply be built into the software, e.g., the piece of information that appears in this location is always a temperature reading in centigrade or, at a different level of granularity, this data set is structured according to Domain Standard A including base types X, Y, and Z where the base types are things like temperature readings in centigrade. This knowledge, easily available within a given domain or a set of closely related research groups, can be built into processing workflows. But outside of that domain or environment the 'local knowledge' approach can begin to fail and more precision in associating data with the information needed to process it is required. This also applies across time as well as domains. What is well known today may be less well-known twenty years hence but age will not necessarily reduce the value of a data set and indeed may increase it.

We are using the term 'type' here as the characterization of data structure at multiple levels of granularity, from individual observations up to and including large data sets. Optimizing the interactions among all of the producers and consumers of digital data requires that those types be defined and permanently associated with the data they describe. Further, the utility of those types requires that they be standardized, unique, and discoverable. The goal of this working group would be to address these issues through evaluation of use cases, existing efforts, and potential infrastructural solutions, including the development of one or more type registries.

Description and Use of Types

Simply listing and describing types in human readable form, say in one or more open access wikis, is certainly better than nothing, but full realization of the potential of types in automated data processing requires a common form of machine readable description of types, i.e., a data model and common expression of that data model. This would not only aid in discoverability but also in the analysis of relations among types and evaluation of overlap and duplication as well as possible bootstrapping of data processing in some cases.

Types will be at different levels of granularity, e.g., individual observation, a set of observations composed into a time series, a set of time series describing a complex phenomenon, and so forth. The ease of composing lower level, or base, types into more complex composite types would be an advantage of a well-managed type system.

An immediate and compelling use case for a managed system of types comes directly out of persistent identifiers (PIDs) for data sets. Accessing a piece of data via a PID, either as a direct reference or as the result of a search, requires resolving the identifier to get the information needed to access the data. This information must be understandable by the client, whether that client is a human or a machine, in order for the client to act on it. For a machine, it must be explicitly typed. A type registry for PID information types would appear to be an early requirement for coherent management of scientific data.

Finally, assigning PIDs to types would aid in their management and use. All of the arguments for using persistent identifiers for important digital information that must remain accessible over long periods of time will apply equally well to whatever form of records are kept for data types.

Type Registries

The set of types used in the management and processing of scientific data must themselves be well managed. Types must be unique and precisely defined in order to be reusable and composable. Creating one or more type registries with common and open interfaces appears to be the best way to accomplish this.

Such registries can add value well beyond accurate description, however, by adding two additional attributes. The first is the source or the authority for the type. Whose idea was this? If further explanation is needed or creation of a new version would be useful, who should be contacted? Secondly, are there services or software available for processing data of the given type? This information could be precisely defined to allow automated processing if the service is available on demand, e.g., if data of type X is sent to service Y the result will be new data of type Z. Such a service registry could be combined with a type registry or exist separately connected by the identifier for the type.

A single universal type registry seems unlikely, if only for organizational reasons. One can envision organizations that would require unfettered control of their own typing mechanisms while allowing some level of federation with others. This would require a level of interoperability, presumably through agreed-upon interface mechanisms as well as agreement on data models and uniqueness. This approach would also raise issues of validation and verification of the federates within a federated set of type registries. This is a role that could perhaps be taken on by the RDA. Finally, federation also raises the issue of interoperation with existing typing efforts.

WG Charter

The Data Type Registries Working Group will

- Compile a set of use cases for data type use and management
- Identify and distinguish among existing ‘type registry’ efforts and their potential interaction with this group
- Formulate a data model and expression for types

- Design a functional specification for type registries

- Propose a federation strategy among multiple type registries at both the technical and organizational levels

Value Proposition

Precise typing of data sets and collections, combined with one or more registries that define those types in a standard fashion, would benefit every sector of data management, especially interoperability and reuse. This WG would not attempt to define the methods of association of data and type but would provide a standard approach for registering and discovering types as well adding value to their use through a standard approach to defining types and for linking types to services.

At least one organization involved in the WG, CNRI, has explicit plans and funding to build a type registry and will use the results of this WG to inform that effort. Further, at least one organization already heavily involved in data management, the International DOI Foundation (IDF) has expressed interest in using and perhaps supporting one instance of a type registry to use in combination with the association of persistent ids and types.

Engagement with Existing Work in the Area

The term 'registry' has many different connotations and meanings across various information management activities and domains. The IANA Mime type registry is a clear example of an existing effort that this group needs to recognize and account for in the context of its goals. At least one of the goals suggested in the Background section above, pointers to relevant services, is not covered by the MIME type registry and it seems unlikely to be so in the future. But MIME types are ubiquitous and well understood, so how would a type registry of the kind envisioned here interact with the IANA registry? Numerous other format registries and service registries have come and gone over the years. This WG should examine those, both successes and failures, to distinguish what is useful and what is and how to interact with those registries and communities that address these issues.

The WG is aware of two communities which have an immediate need for types associated with PIDs – the EUDAT project, represented in this area by EPIC, and the world of Handle System users, specifically the IDF. Both of these groups are represented on this WG and their requirements would be gathered as part of the use case analysis. In addition we envision considerable interaction between this proposed WG and the proposed RDA WG on PID Information Types.

Action Plan

The WG will produce documentation in the form of a set of requirements and a data model for defining data types, and a set of functional requirements for type registries, including federation across type registries. In addition, at least one prototype registry will be built corresponding to that set of requirements. CNRI, represented by one of the Co-Chairs and several members of the WG, has been funded by the Sloan Foundation to build a type registry and anticipates that this WG will strongly inform that activity. That funded started in December of 2012, runs for 18 months, and has as one of its primary deliverables an open source turnkey registry useful across a variety of information management tasks and having a type registry as one use case. Further we anticipate use of that registry by a number of communities also represented in the RDA, including EUDAT and IDF.

Work Plan

The WG will use the RDA forum, email lists, and virtual meetings as required to advance the discussion. CNRI will provide the computing facilities for the prototype type registry and will publish the documentation, using RDA facilities as appropriate and as they come on stream. We anticipate the following approximate timeline:

12/2012 to 2/2013

- Iteration on the Case Statement
- Define scope of effort
- Initial set of virtual meetings to socialize the project

3/2013 – 9/2013

- Gathering use cases
- Investigating other work in the area
- First drafts of data model and functional specs for a type registry

10/2013 – 12/2013

- Refine data model and functional specs
- Deploy initial prototype

1/2014 – 5/2014

- Finalize data model and functional specs
- Deploy functional type registry for PID types
- Release turnkey registry conforming to functional specs

Appendix A: Current WG Membership