PRINTER-FRIENDLY FORMAT Return to Article

D-Lib Magazine

January/February 2014
Volume 20, Number 1/2

Data Type Registries: A Research Data Alliance Working Group

Daan Broeder
Max Planck Institute for Psycholinguistics
daan.broeder@mpi.nl

Laurence Lannom
Corporation for National Research Initiatives
llannom@cnri.reston.va.us

doi:10.1045/january2014-broeder

Abstract

Automated processing of large amounts of scientific data, especially across domains, requires that the data can be selected and parsed without human intervention. Precise characterization of that data, as in typing, is needed once the processing goes beyond the realm of domain specific or local research group assumptions. The Research Data Alliance (RDA) Data Type Registries Working Group (DTR-WG) was assembled to address this issue through the creation of a Data Type Registry methodology, data model, and prototype. The WG was approved by the RDA Council during March of 2013 and will complete its work in mid-2014, in between the third and fourth RDA Plenaries.

Problem Statement

Automated selection and processing of large amounts of scientific data, especially across domains when domain specific tools cannot be used, requires that the data can be parsed and interpreted 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 approach of implicit knowledge available in domain specific tools 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.

Description and Use of Types

We are using the term 'type' here as the characterization of data structure at multiple levels of granularity, from individual data points 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 dynamics of research methodologies and new software tools requires that new types should be easily added to the registries to support new software. The goal of the Research Data Alliance (RDA) Data Type Registries Working Group (DTR-WG) is to address these issues through evaluation of use cases, existing efforts, and the development of an approach to identifying, defining, and making available a set of useful data types via one or more interoperable Data Type Registries.

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 and reuse 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. This we acknowledge is not only a technical challenge but also a social one. But here our position is that we want to offer mechanisms enabling a variety of policies and multiple registry authorities.

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 type registries within the federation. 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.

Figure 1 below shows one possible workflow for use of a Data Type Registries.

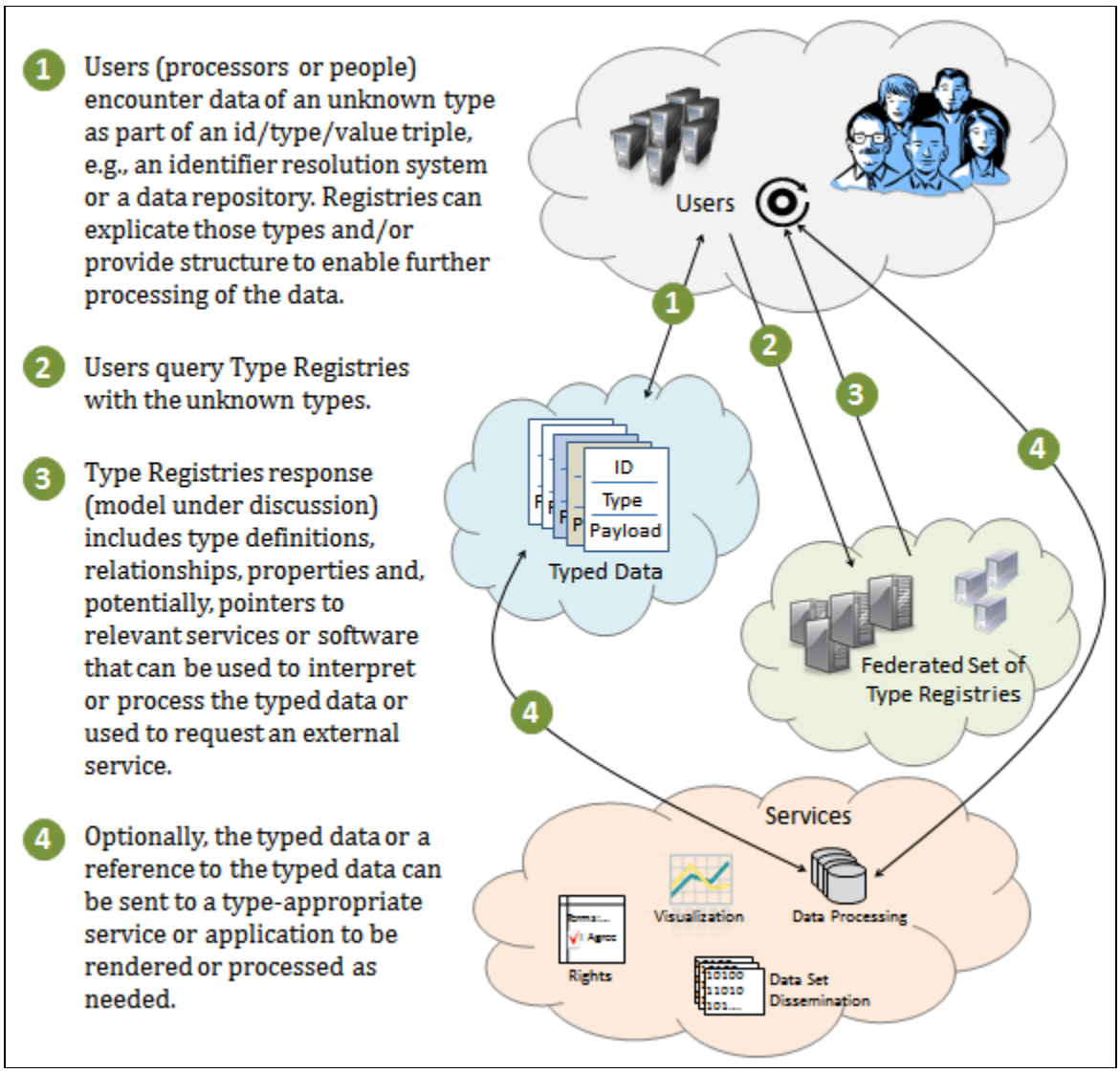

Figure 1

Two alternative uses of such a data type ecology include searching for data by type after using a registry to discover types of use to you and using types as a short-cut for dependent services to figure out if a given data object has what is needed for processing.

Data Model

For a system of data type registries to be useful and interoperable, the records describing types must themselves be well structured and use a standardized set of elements describing the types. Our initial proposal for such a data model, excluding the related services described above, is shown in Figure 2.

Element	Cardinality (min, max)	Notes
ID	(1,1)	A unique, persistent identifier. Assigned by a type registry
Human Description	(1,*)	Description in English mandatory. Descriptions in other languages as needed
Provenance	(1,1)	Who created it, when, etc.
Properties	(0,*)	Properties that describe data. Aka predicates. For example, a weather dataset contains time, location, and temperature properties
Encoding Information	(0,*)	File-formats (mime-types), etc.
Semantic Information	(0,*)	OWL, KIF, etc.
Service Information	(0,*)	WSDL, WADL, APIs, etc.

Figure 2

Use Cases

The DTR-WG effort was informed and is steered by a number of use cases that range from actual description of the types of research data objects within a specific research community to classifications of resource access conditions and classification of object references, which can have a much more general application. This spread from the (community) specific to the possibly very general helps to keep the DTR infrastructure design flexible with respect to community specificity and authoritativeness of the registry information. Currently we have the following use cases:

1. Broad functional classification. In this use case a closed vocabulary of high-level functional classifications for objects kept in repositories and referenced by URIs or PIDs is maintained in a DTR. Examples are that the referenced object can be a primary data object, structured metadata, a general repository-landing page etc. Software can use this information to flexibly consume the primary data, search for the primary data via the metadata or skip this reference because the primary data is not easy available.
2. Data Object Types in the Deep Carbon Observatory Data Management System. This is a community specific use case from the Deep Carbon Observatory (DCO) that seeks to have highly specific descriptions for the types of all objects in the DCO Data Management System (DCOMS). Amongst other uses it will allow the creation of templates for data entry that can map the raw data into a structured object. Since the DCOMS uses the Handle System for referring to objects, the DT identifier will be an associated record with the Handle pointing to the object.
3. Data-Cite access conditions. For data sets that are referenced by Data-Cite issued DOIs, different access conditions can apply. These classified conditions are registered as a type in a DTR. The specific access condition for a data set can be directly associated with the DOI in a HS record. Note that if in use case (1) a PID is used for the object reference, this is just a different attribute of the data object.
4. Discoverable types for HS records. Currently the records associated with a Handle can be typed, and most Handle clients define actions for some (built-in) record types such as HTTP redirection for the URL type. Having a public accessible Handle record type registry would allow tool developers to find useful existing types, and create new ones and define new actions.

5. Community defined profiles of HS record types. This use case builds on the previous use case (4) but adds to this that sets of such HS record types can be defined as a profile by communities of practice, thus forming a separate type. Tools and services can be made aware of such profiles and provide selective functions that depend on the information made available in the handle records from such a profile.

Engagement with Existing Work in the Area

The term 'Registry' in combination with 'Type' or 'Format' 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 Type Registries 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 International DOI Foundation (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.

Discussion on the RDA Forum has produced some useful suggestions about other communities and efforts that should be considered. These include XBRL, the SCIDIP-ES project, and the DataNet Federation Consortium (DFC) Format Registry.

WG Outputs

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 Alfred P. Sloan Foundation to build a type registry and anticipates that this WG will strongly inform that activity. That funded effort 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.

Adoption and Impact

The true value of RDA will be seen in the adoption of RDA Working Group outputs and in the case of the data type registries effort the outlook is already quite promising. The co-chairs of the group put it together knowing about requirements for such a service in both the EUDAT project and in the Handle/DOI communities in general and firm adoption plans are in place now for those projects. But the real benefit of bringing this project to RDA has become clear with the exposure of the effort to other groups within RDA. Multiple other working groups, including those looking at PID Information Types and Practical Policy have both added to the discussion and have their own adoption plans. This interaction among the Working Groups adds considerable value to the efforts that would not be available to them in isolation. In addition to the formal WGs, RDA includes many individual members representing many different projects and areas of expertise. A number of these have joined

the DTR WG and have added to the discussion and promise to further extend the adoption of data type registries.

Conclusion

To date the Data Type Registries Working Group has compiled a set of use cases for data type use and management and will continue to gather additional cases, has examined similar efforts, and drafted a data model. We anticipate continuing into 2014 to refine the data model, add one or more expressions of that model, field a prototype registry, document the effort including a functional specification and federation strategy, and pursue the adoption of data type registries as a data management tool among multiple communities and individual projects.

The stated goal of RDA is to 'build the social and technical bridges that enable data sharing' and we believe that Data Type Registries can be an important tool for building those bridges.

About the Authors

Daan Broeder works in the The Language Archive (TLA) unit at the Max-Planck Institute for Psycholinguistics in Nijmegen, NL for which he is the deputy-head and CTO. He is responsible for the group developing the core LTA archiving software that is also used by several other organizations and institutes. He is currently involved in several EU infrastructure projects and collaborations on Language Resource management, such as the European CLARIN project and its Dutch national pendant CLARIN NL. In both projects the CMDI metadata infrastructure, which he initiated and coordinated, plays an essential role. He is part of national and international standardization groups on language resources. After first working on the development of signal analysis software packages for phonetic research, he switched to developing support for Language Resource data management. He played a major role in the development of the IMDI metadata infrastructure within a number of EU and national projects that is one of the first domain specific metadata sets for the linguistic domain. Currently, his major research interests are developing sustainable e-infrastructures and tools that will effectively eliminate the institutional and organizational boundaries for linguistic research.

Laurence Lannom is Director of Information Services and Vice President at the Corporation for National Research Initiatives (CNRI), where he works with organizations in both the public and private sectors to develop experimental and pilot applications of advanced networking and information management technologies. His current work is focused on CNRI's Digital Object Architecture, which is based on the concept of the digital object, a uniform approach to representing digital information across computing and application environments, both now and into the future. He is responsible for the development and ongoing evolution of a series of infrastructure components needed to implement the architecture, including the widely used Handle System for identifier resolution. He currently serves as co-Chair of Research Data Alliance, US, and as Editor-in-Chief of D-Lib Magazine.