Data Type Registries (DTR)

Karlsruhe

May 28th 2015

Christophe Blanchi

CNRI

Problem: Implicit Assumptions in Data

- Data sharing requires that data can be parsed, understood, and reused by people and applications other than those that created the data
- How do we do this now?
  - For documents – formats are enough, e.g., PDF, and then the document explains itself to humans
  - This doesn't work well with data – numbers are not self-explanatory
    - What does the number 7 mean in cell B27?
- Data producers may not have explicitly specified certain details in the data: measurement units, coordinate systems, variable names, etc.
- Need a way to precisely characterize those assumptions such that they can be identified by humans and machines that were not closely involved in its creation

Goal of the DTR Effort: Explicate and Share Assumptions using Types and Type Registries

- Evaluate and identify a few assumptions in data that can be codified and shared in order to…
- Produce a functioning Registry system that can easily be evaluated by organizations before adoption
  - Highly configurable for changing scope of captured and shared assumptions depending on the domain or organization
  - Supports several Type record dissemination variations
- Design for allowing federation between multiple Registry instances
- The emphasis is not on
  - Identifying every possible assumption and data characteristic applicable for all domains
  - Technology

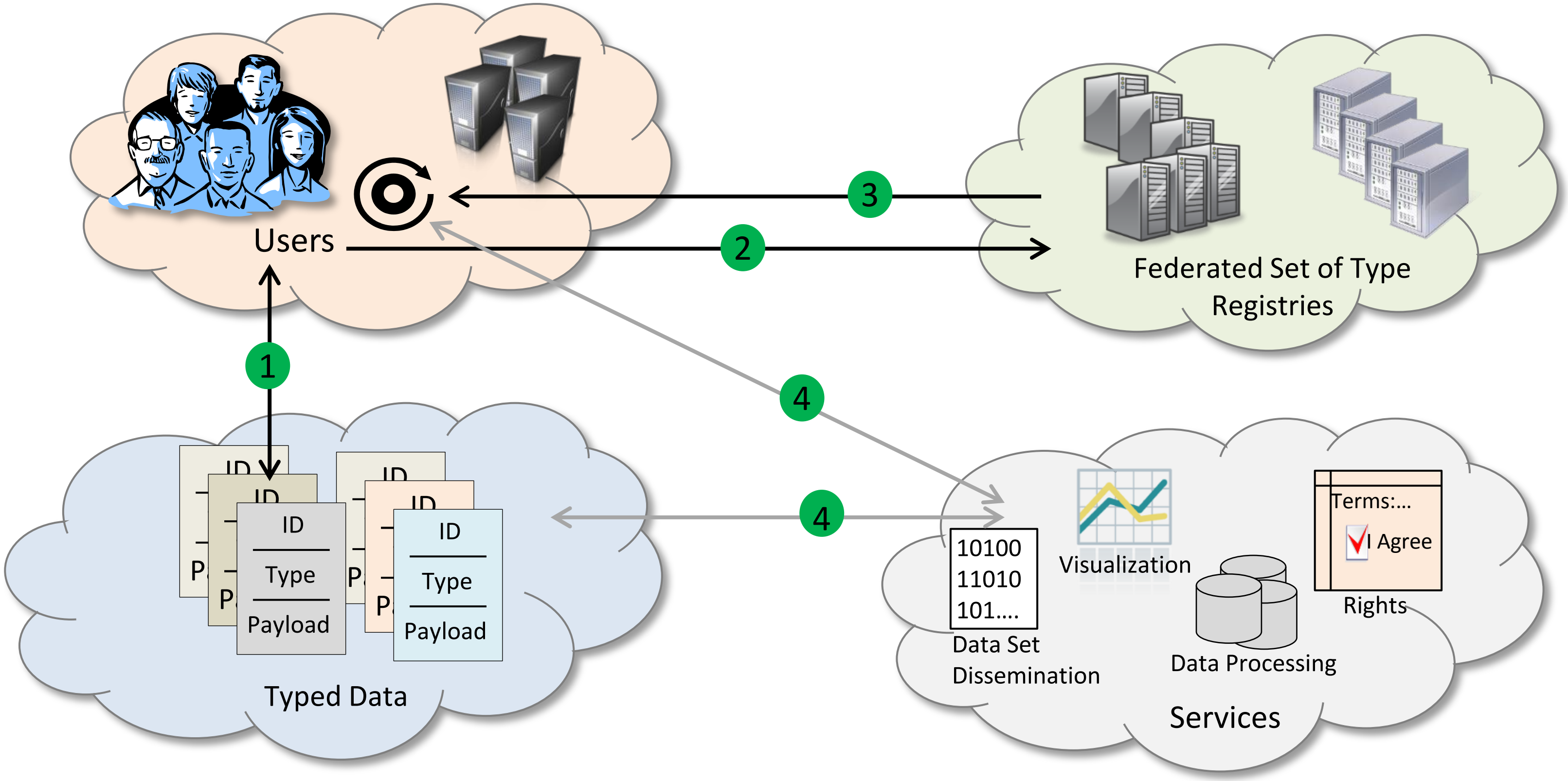
What is a Data Type?

- A unique and resolvable identifier
  - Which resolves to characterization of structures, conventions, semantics, and representations of data
  - Serves as a shortcut for humans and machines to understand and process data
- File formats and mime types have solved the ‘representation’ problem at a ‘unit’ level
- Examples of problems we aim to solve with data types:
  - It is a number in cell A3, but is it temperature? If so, in Celsius?
  - It is a dataset consisting of location, temperature, and time, but what variable names should I look for?
  - Is it all packaged as CSV or NetCDF? And as a single unit or a collection of units?
- Type record structure will continue to evolve – not finished, but functioning

What is a Data Type Registry?

- A low-level infrastructure with wide applicability to record and disseminate type records
  - Not an immediate ROI application
- Assigns unique and resolvable identifiers to type records
- Enforces and validates common data model & expression for interoperation between multiple instances of Registries
- API for machine consumption
- UI for human use

Process Use Case

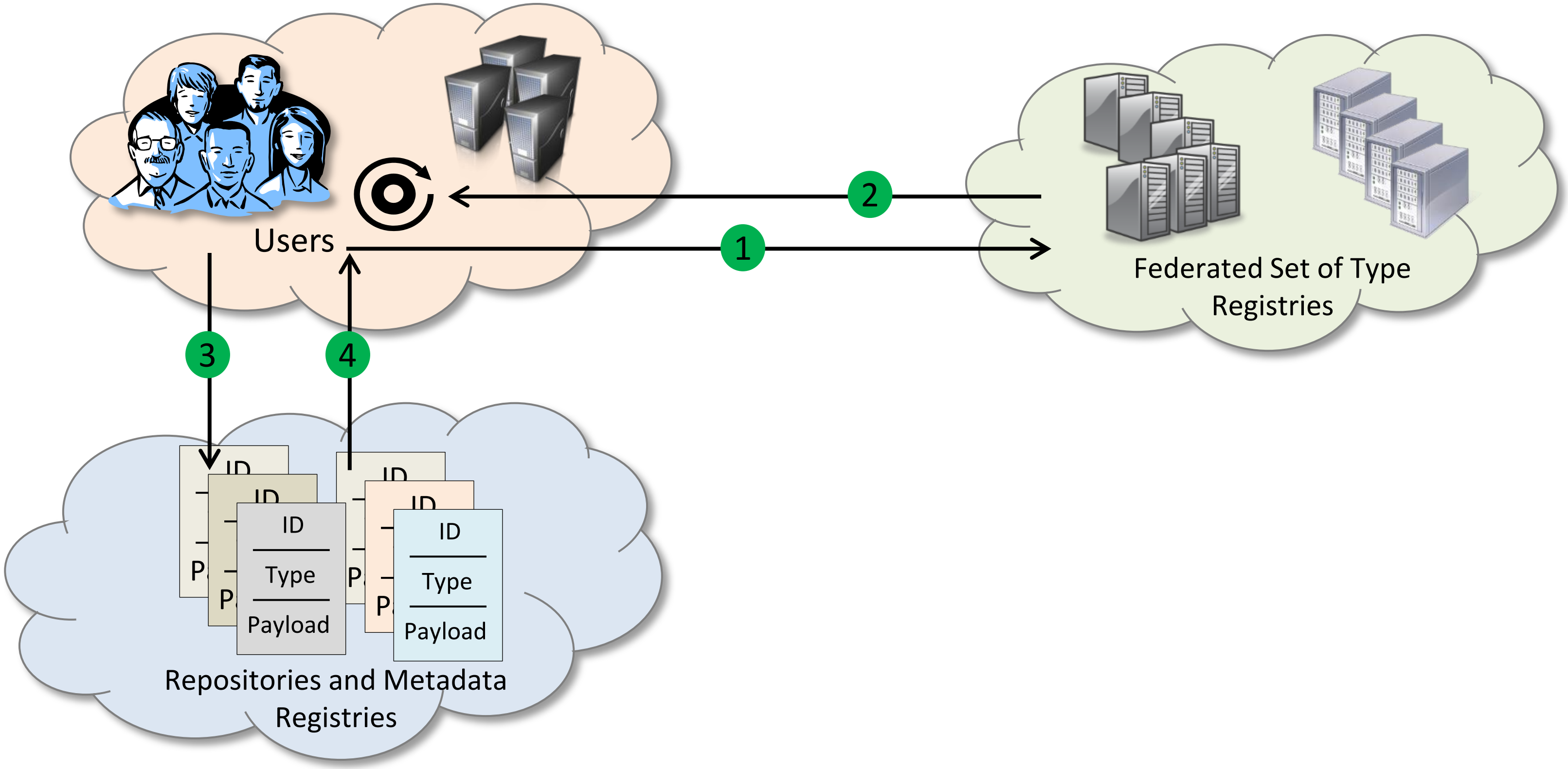


1 Client (process or people) encounters unknown data type.

2 Resolved to Type Registry.

3 Response includes type definitions, relationships, properties, and possibly service pointers. Response can be used locally for processing, or, optionally 4 typed data or reference to typed data can be sent to service provider.

Discovery Use Case


Users
1
2
3
4
Federated Set of Type Registries
ID
Type
Payload
ID
Type
Payload
Repositories and Metadata Registries

1 Clients (process or people) look for types that match their criteria for data. For example, clients may look for types that match certain criteria, e.g., combine location, temperature, and date-time stamp.

2 Type Registry returns matching types.

3 Clients look up in repositories and metadata registries for data sets matching those types.

4 Appropriate typed data is returned.

Type Registry History

- Handle Types – 0.Type/SomeType
  - Good idea, limited applicability
  - Profiles – 'type' the whole set of handle/type/value triples. No traction
- Sloan Grant: 2012 -2014
  - Generic Registry system using Type Registry as a use case
- NSF Grant: 2013 -2014
  - Included support for Type Registry
- Research Data Alliance (RDA) Data Type Registries Working Group
  - One of first two WGs approved at Plenary 1 – March 2013
  - International representation, > 50 members
  - Co-chairs Lannom (CNRI), Broeder (Max Planck Psycholinguistics Institute)
  - Should result in an RDA Recommendation (2015?)
- International DOI Foundation
  - Proposed set of standard types for certain functions, e.g., resolve to license
  - IDF-specific Type Registry

Current State

- A prototype is at: http://typeregistry.org/
- Implementation supports notions of primitives and derived types
- Primitives are fundamental types that we expect humans and software to parse and understand
  - Integer, floating point, boolean value, string, date, timestamp, etc.
- Derived types depend on primitives to describe something complex
  - Stream gauge, Lidar, Spatial bounding box, etc.
- Registered types are assigned unique identifiers

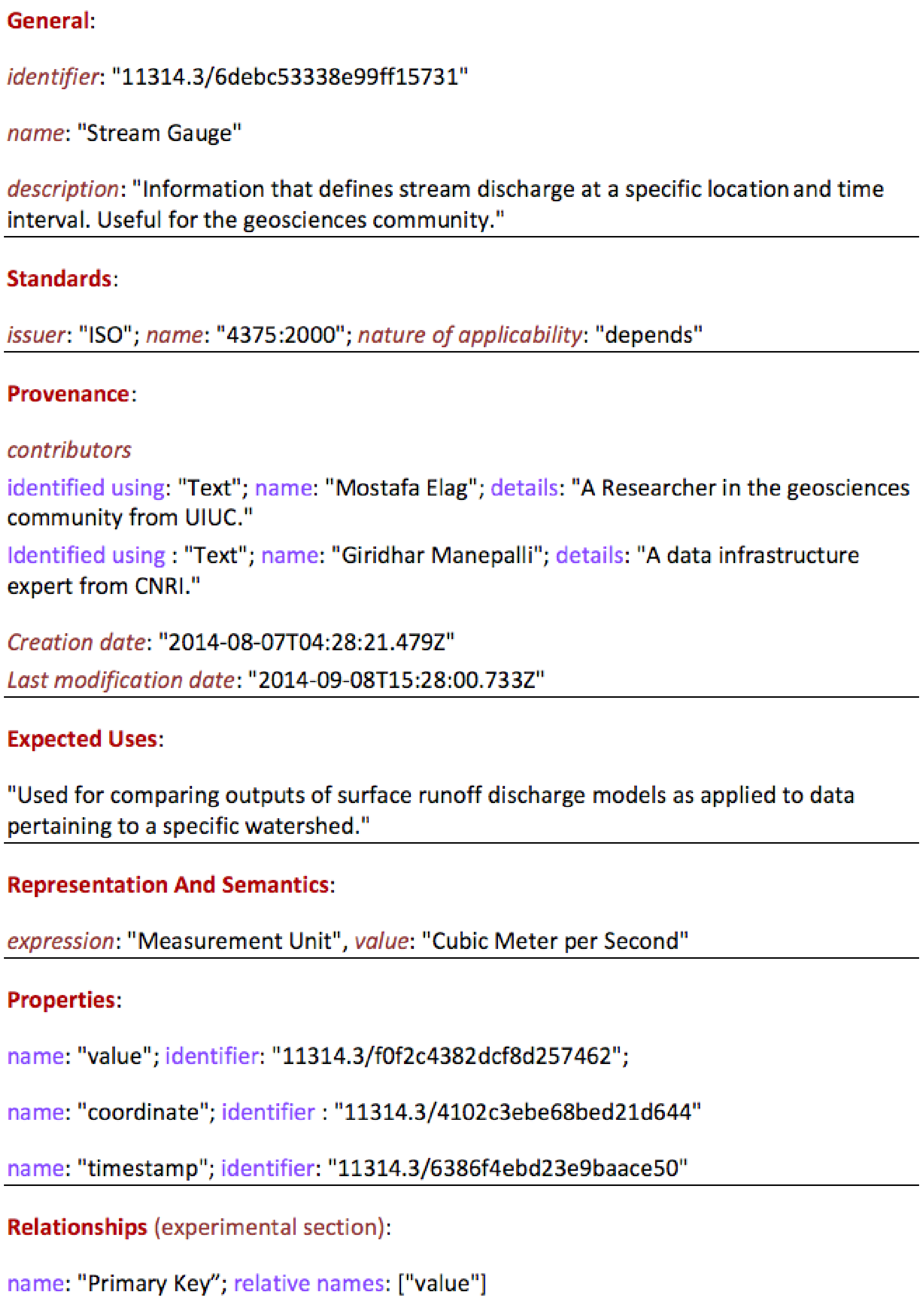
What Has the DTR WG Accomplished?

- Confirmation that detailed and precise data typing is a key consideration in data sharing and reuse and that a federated registry system for such types is highly desirable and needs to accommodate each community’s own requirements
- Deployment of a prototype registry implementing one potential data model, against which various use cases can be tested
- Involvement of multiple ongoing scientific data management efforts, across a variety of domains, in actively planning for and testing the use of data types and associated registries in their data management efforts
- Integration with one additional RDA WG (Persistent Identifier Types) and at least one Interest Group (RDA/CODATA Materials Data, Infrastructure & Interoperability IG)
- Development of a set of questions that require further consideration before a detailed recommendation on data typing can be issued

What are the High Level Data Type Registry Requirements?

- Every type in a data type registry must be identified with a resolvable persistent identifier
- Types should reference related standards and recommendations in order to leverage existing efforts
- Primitive types should be established and used, when possible, in the construction of more complex types
- A common API should be available across all type registries
- Type registries should be federated such that a single service can search across all known registries
- Type registries should include or enable referencing related services based on types
- The establishment of a data type registry for any community should be subject only to the needs and requirements of that community, i.e., there should be no higher level governance beyond the maintenance of whatever standards and processes are needed for effective federation across type registries

Data Type Example



General:

identifier: "11314.3/6debc53338e99ff15731"

name: "Stream Gauge"

description: "Information that defines stream discharge at a specific location and time interval. Useful for the geosciences community."

Standards:

issuer: "ISO"; *name*: "4375:2000"; *nature of applicability*: "depends"

Provenance:

contributors

identified using: "Text"; name: "Mostafa Elag"; details: "A Researcher in the geosciences community from UIUC."

Identified using : "Text"; name: "Giridhar Manepalli"; details: "A data infrastructure expert from CNRI."

Creation date: "2014-08-07T04:28:21.479Z"

Last modification date: "2014-09-08T15:28:00.733Z"

Expected Uses:

"Used for comparing outputs of surface runoff discharge models as applied to data pertaining to a specific watershed."

Representation And Semantics:

expression: "Measurement Unit", *value*: "Cubic Meter per Second"

Properties:

name: "value"; identifier: "11314.3/f0f2c4382dcf8d257462";

name: "coordinate"; identifier : "11314.3/4102c3ebe68bed21d644"

name: "timestamp"; identifier: "11314.3/6386f4ebd23e9baace50"

Relationships (experimental section):

name: "Primary Key"; relative names: ["value"]

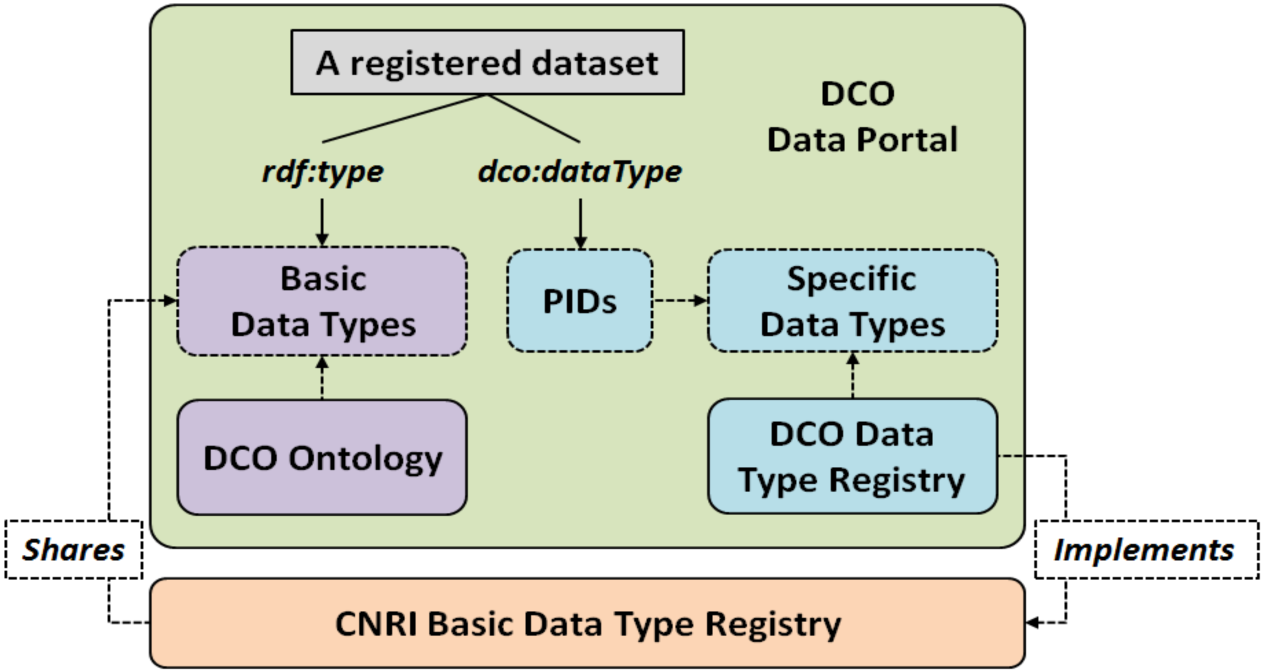
Deep Carbon Observatory (DCO)

- A multidisciplinary, international initiative dedicated to achieving a transformational understanding of Earth's deep carbon cycle
- DCO Science Network consists of more than 1700 scientists from 400 organizations and 40 countries
- A conceptual model of the interplay between data, people, publication, instruments, models, organizations, repositories, etc.
- Identify, annotate and link all key entities, agents and activities
- A repository for datasets and associated metadata
- Data and metadata visualization for dissemination of information
- Collaboration tools for scientific efforts
- An integrated portal for diverse content and applications

DCO Plans for DTR and PIT

- DCO Data Portal provides the digital object registration process for DCO Community members, which includes
  - DCO-ID handle generation based on the global Handle System
  - metadata collection for each registered object.
- Datasets in the DCO community cover various formats and topics in Earth and space sciences.
- Goal: given a dataset identifier, discover detailed information about the structure(s) within that dataset, and act accordingly.
- PIT provides a general model for connecting identifiers and types
- DTR provides a registry for explicating types
- Facilitate norms of behavior relevant to data curation and re-use.

DCO Data Portal and DTR



- DCO basic types held as primitives in the ‘base’ DTR
- DCO –specific DTR extends primitives

(Figure courtesy of the DCO Data Science team at Rensselaer Polytechnic Institute.)

Materials Genome Initiative (MGI)

- Materials Genome Initiative intended to enable discovery, development, manufacturing, and deployment of advanced materials at least twice as fast as possible today, at a fraction of the cost
- At the heart of MGI is the Materials Innovation Infrastructure [MII], a framework of seamlessly integrated advanced modeling, data, and experimental tools
- MGI aims to link together networks of scientists spanning academia, National and Federal laboratories, and industry to more effectively share the information that underpins new material discovery and product development, and enables technological leaps
- NIST is one of the six Federal agencies that comprise the Subcommittee on the Materials Genome Initiative

MGI (Kent State) Plans for DTR & PIT

- Focus on a Use Case to develop an improved turbine blade with the capability to withstand higher temperatures for improved fuel efficiency in the aerospace industry
- Test the front-end of the RDA Data Type Registry WG's product in consultation with the RDA PID Information Types WG
- Work closely with NIST to obtain relevant small and large datasets, as well as guidance, and feedback.
- The proposed 5-month project seeks to identify relevant data types to be connected with front-end applications and services of the data producer required in the Use Case and so enable data consumers to perform analysis through backend applications and services

US Census Bureau

- Conducts various surveys to gather and analyze social, economic, and geographic status in US
- Data gathered from surveys is synthesized before being exposed for outside analysis
- Synthesized data, therefore, comes packed with multiple assumptions made by surveys. Examples of such assumptions are
  - Income dataset of a particular region is only about minorities
  - Home sales dataset considered only homes sold by primary residents
- Actual assumptions are much more complex, nuanced, and granular
- Goal: Two fold
  - Create data types to characterize each column of each synthesized dataset at sufficient granularity to enable humans and applications "process values"
  - Codify and represent underlying assumptions within data types so humans and applications can process values "without introducing statistical errors"