WG: Data Type Registries

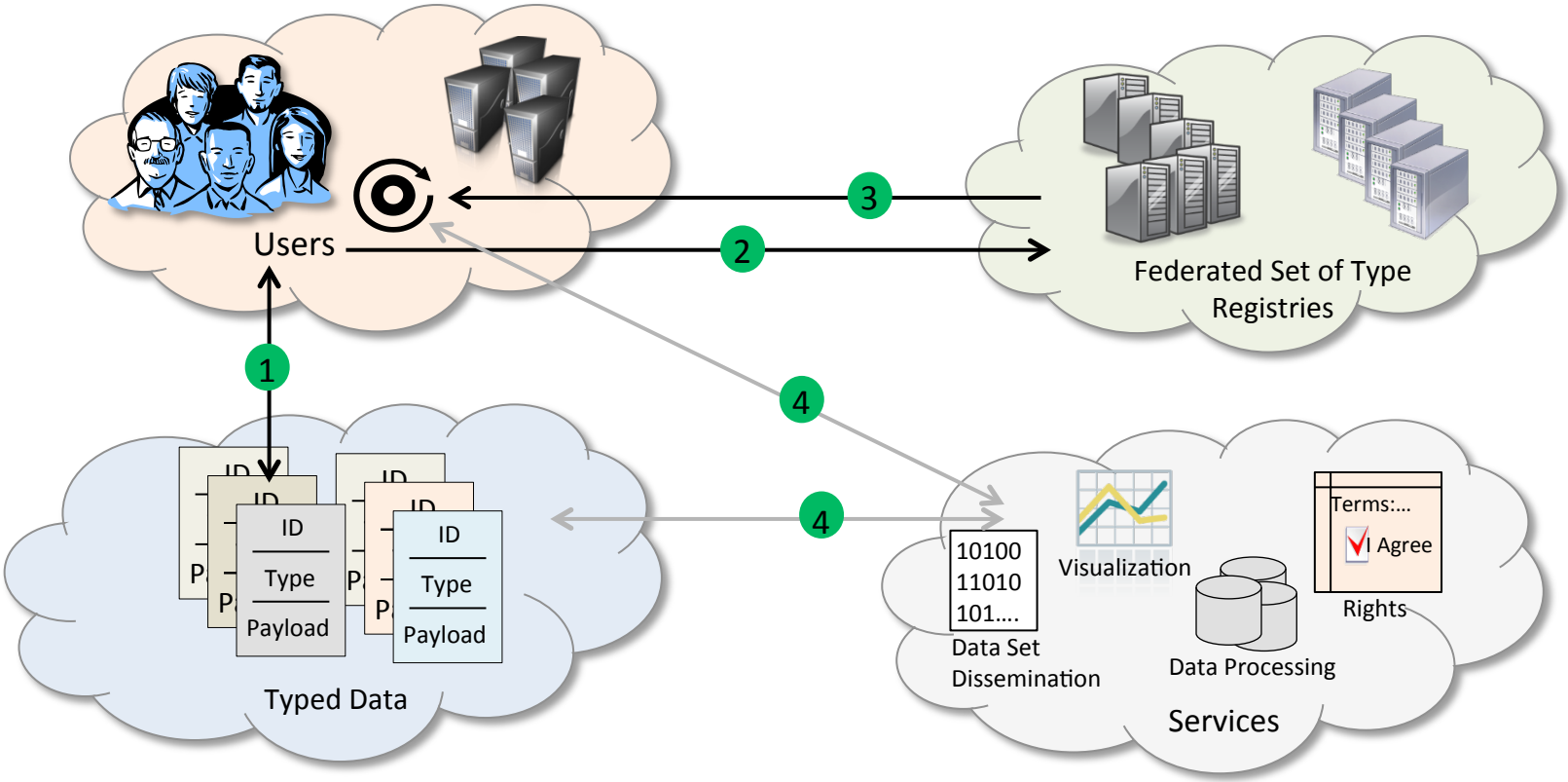
3rd RDA WG Collaboration Meeting
Karlsruhe
11 June 2015

Larry Lannom

Data Type Registries (DTR) WG

- Problem: Data sharing requires that data can be parsed, understood, and reused by people and applications other than those that created the data
  - Current format registries, MIME types, etc., don't quite fill this gap
- Solution: Registry framework allowing easy registration and identification of precise data definitions at multiple levels of granularity that can be referenced from within data sets and/or data set metadata
- Example: Stream Gauge Measurement data type that defines the elements of a set of measurements from sensors in streams and rivers
  - Time stamp in known units
  - Location in known units
  - Volume per time, e.g., cubic meters per second

Data Type Registry (DTR) 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.

Alternative use cases include discovery of data matching a certain type

DTR Prototyping

- Deep Carbon Observatory (DCO)
- Materials Genome Initiative (MGI)
- International Digital Object Identifier (DOI) Foundation
- U.S. Census Bureau
- Additional Groups Planning Prototypes
  - Woods Hole (research cruises)
  - U. Washington (air quality monitoring)
  - EUDAT
  - NIST

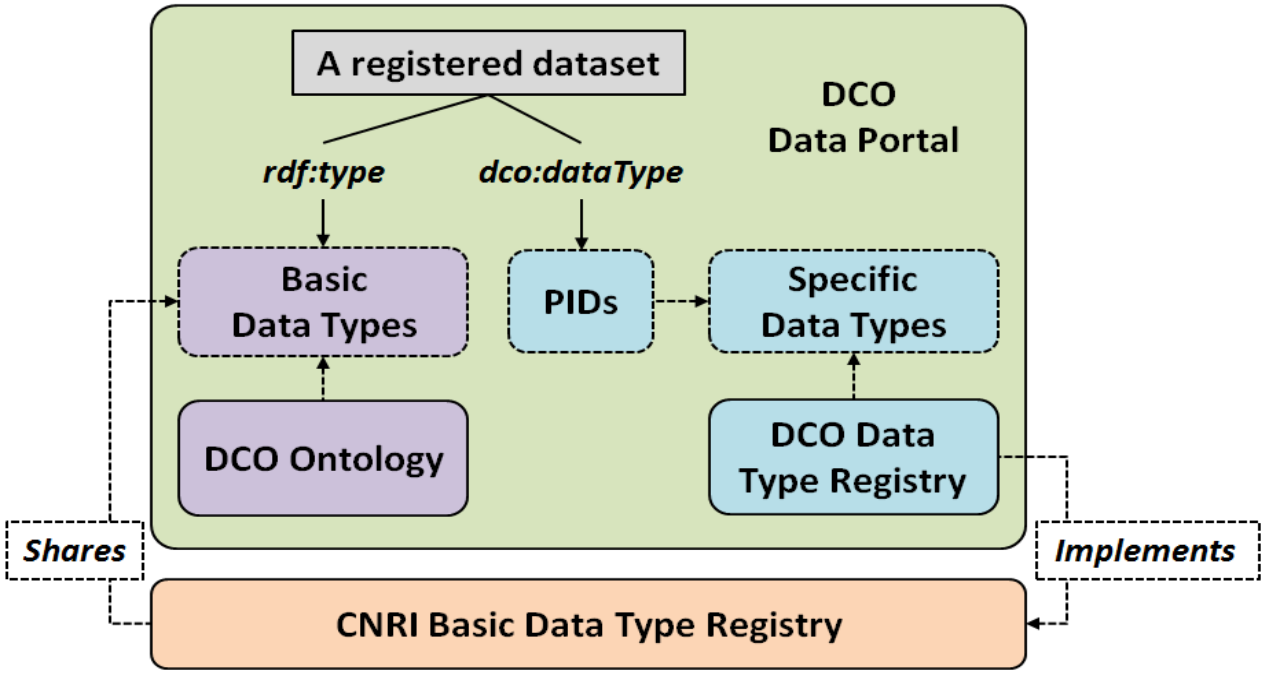
Deep Carbon Observatory (DCO)

- 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

Deep Carbon Observatory (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

Materials Genome Initiative (Kent State) Plans for DTR & PIT

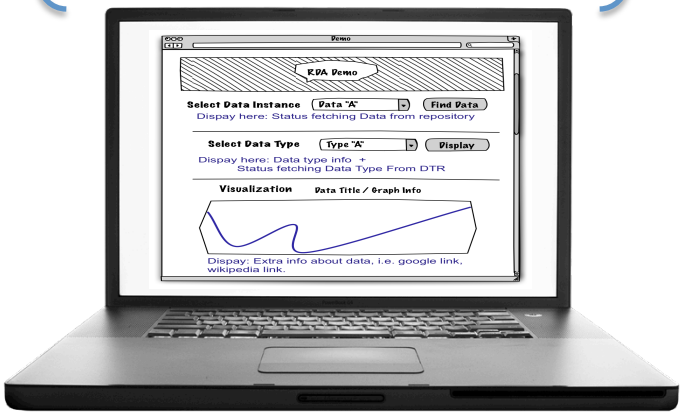
- Focus on a Use Case – chose x-ray diffraction – normalize data sets resulting from multiple proprietary instruments
- Test the front-end of the RDA Data Type Registry WG's product in consultation with the RDA PID Information Types WG
- Work closely with NIST to obtain relevant small and large datasets, 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

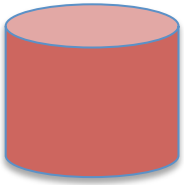
Use Case for MGI/RDA Demo

- Discovery Visualization Tool (DVT) to query resource registry for looking for all data containing Al2O3
- DVT to narrow the results by: *is-a diffractogram datatype of any representation*
- This will result in 11 results conforming to 5 different representation datatypes (gss,dat,cpi,prn,xda)
- DVT to query resource registry for a resource than can convert from representation type (gss,dat,cpi,prn,xda) to canonical 2D figure representation type
- DVT to render generated json representation

Revised Use Case for MGI/RDA Demo

7. Discovered 11 datasets, conforming to 5 data types. Unable to plot any. Searching for conversion resource.

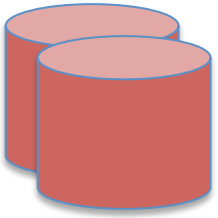




NIST Registry

1. Al_2O_3?

2. Check Resource A and J



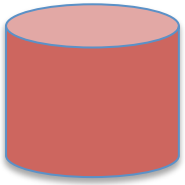
3. Fetch Datasets

4. Datasets Returned

Repositories

5. Fetch Data Types

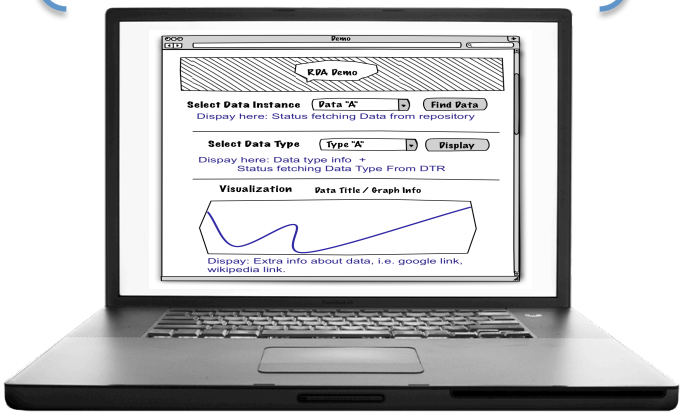
6. Data Types returned



Data Type Registry

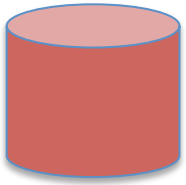
Revised Use Case for MGI/RDA Demo

7. Discovered 11 datasets, conforming to 5 data types. Unable to plot any. Searching for conversion resource.



8. Conversion resources?: 11314.3/368d to 11314.3/74b5 and...

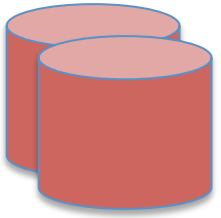
9. Check Resource-M



NIST Registry

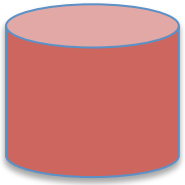
10. Proprietary Format

11. Canonical Format



Resources

12. Visualize



Data Type Registry

Other DTR Prototyping

- International Digital Object Identifier (DOI) Foundation
  - Over 100M resolvable identifiers across publishing, data, and entertainment
  - Each resolution returns one or more type/value pairs
  - Connect services to resolution types
- U.S. Census Bureau
  - Create data types to characterize each column of each synthesized dataset at sufficient granularity to enable humans and applications to “process values”
  - Codify and represent underlying assumptions within data types so humans and applications can process values “without introducing statistical errors”

What Did the DTR WG Accomplish?

- 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
- Multiple Prototyping Efforts Now Underway
- Included in EUDAT call for Pilot Projects
- Data Typing (DT) WG in planning stages