Data Foundations And Terminology (DFT) IG

shared conceptualizations

research data sharing without barriers
rd-alliance.org

DFT IG Breakout Session       P 13 Breakout – 4 April
Co-Chairs DFT IG : Gary Berg-Cross & Raphael Ritz
Overview of Objectives (and Agenda) for P-13

1. Updates and Continue IG discussion –
   Who is completing work and has vocabularies?
   • How do they relate to each other?
   • Tool update
   • Start on a model for FAIR and Open Data concepts
   • Will this extend the RDA vocabulary core?

2. Notes on facilitating community discussion on RDA/group core concepts
   • Systematize the already large body of domain definition work on terms and their meaning
   **Use a rationalized “consensus”** knowledge of domain experts, especially for Internal RDA’s efforts but also external collaboration & coordination.

3. Re-organization of Termed Concepts?

4. Lightweight Metadata/Vocabulary Methods and the HyperModel problem

5. Interest in broader vocabulary effort and Next Steps?
Open & FAIR Digital Data Management is a broader concept

Digital Object Management (registered, digital data)

Organized in a FAIR DO EcoSystem

Where are datasets???

State Info is MD about current state of a DO
Some Vocabulary around P8

- **Collections** and various types of **metadata** were actively discussed:
  - **Descriptive metadata** is a type of metadata that describes a resource for purposes such as discovery and identification and **Contextual Metadata**
  - **Curation** metadata describe who supports a curated resource and its availability.
  - **Data Citation Metadata** is a type of metadata/Administrative metadata that plays the role of citing a dataset in an analogous way that books or journals or a computer, access date, version number, and a persistent identifier or locator. Metadata that maps to DataCite schema or Dublin Core Terms etc.
  - **Domain Metadata** or domain-specific metadata is non-general metadata used to capture domain information as reflected in domain vocabularies and when possible domain-specific metadata should maps to metadata standards used within a scientific domain.
  - **Rich Metadata** describes data with enough accurate and relevant attributes to make it easily findable.
  - **Key Metadata** is information associated with a digital object (or entity) that are required for discovery.
  - **Metadata Catalogue** A type of data catalog (catalogue) used to access information about data

- **Data Transparency; Fair Use; Patent; Copyright Infringement; Rights Statement; etc.**
1. **Administrative Metadata** is a type of Metadata that provides information to help manage a resource, such as when and how data was created, a file type and other technical information, and who can access it.

2. **Authenticity Metadata** is a type of metadata that conveys information needed to link a data object to its original source with integrity. Authenticity is provided by appropriate metadata, within an archive & digital retention and preservation context. It results from verifying that a digital object & its state information has not changed.

3. **Citation Metadata** serves the role of identification and should provide an unambiguous identifier to the data cited, its location, and means of access.

4. **Detailed Metadata** is defined in distinction to simpler or light forms of metadata that provide some basic information about data, such as in Dublin Core, but which can supplement this simple information.

5. **Discovery Metadata** is metadata whose chief role is to discover relevant data.

6. **Extract Descriptive Metadata** works by using a given data type, to access a data type registry and identify a procedure that can be used to parse the data object and then apply a template to extract desired information from the contents of the data object.

7. **Key Metadata** is information associated with a digital object (or entity) that is required for discovery. Thus it is a part of Discovery Metadata.

8. **Objective Metadata** is based on assertions of fact about such things as authorship, date of creation, & version. Broadly they include attributes can be assigned by what is considered an objective and reproducible (perhaps automated) process.

9. **Minimal Metadata** descriptions with very little curation including DO name & PID - only marginally targeted at the role of discovery since there is much better infrastructure to accomplish this.

10. **Payload, Provenance, Rich, Structural System, Topical metadata, FAIR MD?**
Recent Additions/Edits Concepts that RDA is “talking about”

**February-March**

FAIR EcoSystem

Data object type, Certified repository. Data Science; Data-driven science...or inductive science is one where scientists discover new knowledge by systematically processing large volumes or complex collections of data ...

Data fragmentation; Data infrastructure; Open science cloud...

**January**

Research Object; Data sharing culture; “is a data culture includes that as part of it priorities and practices incentives for sharing, guidance via data sharing policy,..."

**Sept- December**

FAIR Metrics Data Publication; (from the NLM data thesaurus.)

Linked Data; Data Catalog; Data usefulness; Data Access Data Discovery Data Sharing (added principles); Metadata schema; Discoverability; Data standards; Preservation; Data Curation

C2CAMP – had many new terms.....

Science Ecosystem, digital object model, tightly associated metadata, Object typing, Mapping services, Global Digital Object Cloud, virtually aggregated digital objects, Solution Space ?????
Out Tool - TeD-T Background and Motivation

• Provide an online tool to coordinate and document community contributions to the development of the RDA DFT Vocabulary

• Software stack: Semantic Media Wiki plus Semantic Forms (think “enhanced wikipedia” software)

• Customizations:
  - Pages have a substructure
  - Various overviews
  - Simple categorization

• Download content in multiple formats

• Referable by PID (vocabulary – terms – definitions)

• Run by the Max Planck Computing and Data Facility (MPCDF) in Germany
Welcome to TeD-T, the Term Definition Tool of the Data Foundation and Terminology Interest Group (DFT IG) of the Research Data Alliance (RDA).

The tool was originally put in place to support the DFT Working Group (2013-2015). The DFT WG task was to describe a basic, abstract data organization model which can be used to derive a reference data terminology that can be used across communities and stakeholders to better synchronize conceptualization, to enable better understanding within and between communities and finally to stimulate tool building, such as for data services, supportive of the basic model’s use. We assume that this abstract data organization model will focus on common building blocks and their characteristics, along with relevant protocols.

After termination of the Working Group the tool continues to support the then established DFT Interest
Structure of a page

- Definition
- Explanation
- Examples
- References
- Scope
- Status

Possibly multiple times

Plus categorization
## Repository Registry

| **Definition** | A repository Registry is a type of registry that collect useful information about repositories for human consumption in order that depositors and users can easily find where to go to for their data needs. |
| **Explanation** | |
| **Examples** | re3data is an example |
| **Scope** | RDA Data Fabric Interest Group |
| **Status** | New |

**Category:** Infrastructure
Alternative Organization Of Termed Concepts

- Data
- MetaData
- Governance
- Policy
- Infrastructure
- Standards
- Semantics
- Functionality
- Services
A FAIR DO is a sub-type of a DO operating according to FAIR principles and managed as part of a FAIR ecosystem/e-infrastructure.

At the core of DO infrastructures is the DO Interface Protocol (DOIP) that defines an interoperability layer between repositories and registries of Digital Objects.
General Vocabulary Methods

There is value in Semantic Annotation & attached processable metadata to data

Simple example such as semantic trajectory
Methods exist to add lightweight semantics for some topics, but in many they aren't routine

- Richer Metadata for Location and Topic/Subject are good areas
  - For example there are many types of located parts-whole relations
- Ontologies can help a bit but even Standard ontologies may need to be adapted to the complex ideas implied in new vocabularies
  - And there are issue with standardizing definitions
- With so many sources for metadata information & Big Data we have a Hypermodel problem:
  - how to connect local ideas into a knowledge system?
Backup Slides
Other Data Management Vocabularies
opportunities for collaboration, coordination, and de-duplication of effort.

Despite decades of intensive work on controlled vocabularies (standardized sets of terms) problems remain with definitions that are central to RDM.

The important need for clear definitions of RDM terms is widely recognized RDA’s Data Foundations and Terminology (DFT) WG is one of the earlier initiatives.

Other important efforts include:

• Science Europe Data Glossary;

• Data Documentation Initiative (DDI); and

• Research Data Canada (RDC)/CASRAI RDM pilot
  • evolved into a new International Research Data Management glossary (IRiDiuM) supported by RDC, CASRAI, and CODATA.

• Update from Big Data at NIST and IEEE workshop...