Overview of Data Foundations & Terminology (DFT) WG/IG
https://rd-alliance.org/groups/data-foundations-and-terminology-ig.html
Gary Berg-Cross (RDA US Advisory Committee)
First Products of the Research Data Alliance (RDA): Foundations

iassist 2015
Bridging the Data Divide: Data in the International Context
MINNEAPOLIS, MN
June 2 - 5

research data sharing without barriers
rd-alliance.org
RDA Scope: We are Talking about Data Management & Infrastructure: with a huge range of concepts and terms that describe the entities that make up “CyberInfrastructure” & circumscribe the multiple relationships they can have with each other.

Sensor networks, scientific instruments, MRECs, streaming data, modeling, computation, simulation, data collections, file systems, databases, government data, long tail, etc.

Acquisition
Storage
Preservation
Interoperability
Access
Services

Metadata
Identifiers
Curation
Policy
Discovery
Analysis
Tools

Discovery, Insight, Re-use

Challenges: heterogeneous categorizations and varied methods for organizing information.
Shared Vocabularies are Needed for Understanding

We build vocabularies to capture an understanding of knowledge in order to:

• facilitate **communication** across disciplines,
  • Research data communities are helped by standardized data vocabularies reflecting the same definition for the same terms (data asset, data object, metadata....)

• **share/exchange data** and reuse it or

• enable **collaboration**

The recent trend is to leverage agreed upon conceptualizations & augment it with a formalized, digital representational of this knowledge allowing some degree of automated processing.

Knowledge is a socially distributed phenomena.
RDA DFT Work Group

• DFT WG was designed to help the RDA community find common building blocks, describing their properties and defining data process protocols related to them.
  • Co-chairs: Gary Berg-Cross, Peter Wittenburg, Raphael Ritz

• Anticipated benefits:
  • Conversations/interactions can be more meaningful
    • people aren't talking past each other but share a space of understanding with a domain community.
  • Helps adoption of common data sharing practices and interoperation
  • Help avoid duplication of effort.

Gardens of RDA Projects
We need.....

Infrastructure,
Metadata, services,
repositories

Foundation
Concepts for ...

Leverage various models like OAIS & get community agreement

Existing Reference Models/vocs

Partners

WGs
Something to Keep in Mind

How standards proliferate:
(See: A/C chargers, character encodings, instant messaging, etc.)

Situation: There are 14 competing standards.

14?! Ridiculous! We need to develop one universal standard that covers everyone's use cases. Yeah!

Soon: Situation: There are 15 competing standards.
Background - DFT WG Activities & Accomplishments

• One of the first RDA WGs

• Goal: Describe a basic, abstract (but clear) data organization model that systemizes the already large body of definition work on data management terms, especially as involved in RDA’s efforts.

• Drafted 4 related Model Documents as part of work:
  1. Data Models 1: Overview – 20 + models
  2. Data Models 2: Analysis & Synthesis
  3. Data Models 3: Term Snapshot of core terms of interest to RDA groups
  4. Data Models 4: Use Cases- Work with other RDA WGs on use cases to illustrate data concepts & drive vocabulary development
  5. Developed Semantic Media Wiki Term Definition Tool (Ted-T) to capture initial list of terms and definitions for discussions (see http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page)

• Participated in Adoption Day - Common Language Resources and Technology Infrastructure Adopting DFT, DataFed.net, CLARIN etc.
Policy defines this?
1. What elements are in a PID record?
2. How to point to a metadata record?
3. What is in a metadata record at registration?
4. What is replication with identical vs. different bit-streams that may store additional attributes?

We have policy for a minimum metadata record?

The rest of data management the lifecycle and data publication?

Other WGs.....
Vocabulary can be Built out in Stages —adapted example from P2

- **Starter Set**: 9-12 months P4-P5
  - Firm up products

- **Core**: 6-9 months P3-P4
  - Get consensus & document well

- **WG Scope**: At 3 months P2

- **Incremental RDA Scope**: 18 months + IG work
  - Expand vocabulary
**Overview of Term Development**

**Starter areas and items:**
- Persistent Identifiers (PIDs and types)
- Digital Object - Data Object
- Collection - Data Set - Aggregation
- Repository (Registries and related Policies)

**Digital Object**

A digital object is composed of a structured sequence of bits/bytes. As an object it is named. This bit sequence can be identified & accessed by a unique and persistent identifier or by use of referencing attributes describing its properties.

**Data and Digital Objects/Entities**

- Data Object
- Metadata Object
- Representation Object
- Information Object
- Active Data
- Service Object
- Digital Object of bits and bytes
- Workflow Object

---

- **Persistent Identifiers (PIDs and types)**
- **Digital Object - Data Object**
- **Collection - Data Set - Aggregation**
- **Repository (Registries and related Policies)**

---

- Defs were organized & prepared for review

---

- Term Definition Tool prototyped and developed at Rechenzentrum Garching (RZG) der Max-Planck-Gesellschaft

---

- Analysis and Revision Process

---

- Models were organized & reviewed

---

- Term Definition Tool prototyped and developed at Rechenzentrum Garching (RZG) der Max-Planck-Gesellschaft
More Terms and Initial definitions are in TeD-T

smw-rda.esc.rzg.mpg.de/index.php/Special:AllPages
Examples of Guiding Use Cases

EUDAT Replication Service

Checksums added to assure that the bitstrings are preserved during replication

Collection building across repositories is essential for running analysis. PhD students, for example, create their collection, which consists of a MD object having a PID and storing many PIDs to refer to its components. Such a collection is an aggregation, but has an identity, can be cited, etc.
Data streams become DO by adding metadata and IDs as they are registered.

No, collection is a special type of aggregation.
Broader View of Data Processes
Much to define and richer ways to define it

- Boarder WG and Data Concept View
  - Cross-group coordinated with several RDA WGs, including the Data Fabric view data concepts and relations.
    - Coordination ongoing as part of an IG.
    - Potentially all groups could be engaged in this IG and we with them
  - PP along with MIG has expressed an interest in more formalized definitions that can be processed by computer and the Ted-T tool may be capable of doing this or at least demonstrating its feasibility.
    - Automated policy driven definitions – example goal
      - To identify typical application scenarios for policies such as replication, preservation etc.
    - Selective examples from metadata
      - Is Web view of data important to include?

- Better Vocabulary Development
  - Taxonomies, relations, attributes

Much more work and discussion would be useful such as with the PP WG and its terminology that was only briefly sketched out without full definitions.
Practical Policy WG area examples

- Contextual metadata extraction
- Data access control
- Data backup
- Data format control
- Data retention
- Disposition
- Integrity (including replication)
- Notification

### Contextual metadata extraction policies

This policy area focuses on metadata associated with files and collections.

The creation of **provenance** and **descriptive** metadata defines a **context** for interpreting the relevance of files in a collection.

Depending upon the data source, there are multiple ways to provide metadata—**some automatable**:

- Extract metadata from an associated document. An example is the medical imaging format DICOM.
- Extract metadata from a structured document which includes **internal metadata**.
  - Examples are FITS for astronomy, netCDF, and HDF.
- Extract metadata by **parsing** patterns within the text within a document.
- **Identify a feature** present within a file and **label** the file with the location of the feature that is present within the file.

<table>
<thead>
<tr>
<th>Extract metadata</th>
<th>Attribute_name</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Attribute_value</td>
</tr>
<tr>
<td></td>
<td>Attribute_unit</td>
</tr>
<tr>
<td></td>
<td>Source_file</td>
</tr>
<tr>
<td></td>
<td>Source_collection</td>
</tr>
</tbody>
</table>
Lessons Learned and Follow Up

• Difficult to get consensus on the scope a common vocabulary with detailed definitions.
  • More model and vocabulary identification than integrated definitions
  • Continued discussions with communities about our results

• Based on interest a DFT IG was formed to continue efforts.

• As part of an IG a broader plan for long-term maintenance & defs upgrades

• A plan for term tool (TED-T) maintenance
  • updates for DFT terms and other WGs.
Objectives for IG

1. Continue DFT discussion and leverage existing work and approach but improve both
   • We are expecting considerable discussion of new requirements coming out of groups just completed or nearing completion, but also support as part of adoption.

2. Focus on facilitating community discussion on core concepts. The current synthesis model can be expected to finalize and stabilize the effort for subsequent use.

3. Facilitate definition development
   • Potential adopters were encouraged at P5 to provide feedback on additional use case scenarios to illustrate what areas of work they plan on using the models and vocabulary for.
   • We now have virtual meetings between P5 and P6.
Backup Slides
5 Major document products of the DFT WG:

- **DFT 1: Model Overview**  An annotated collection of data organization & management models that represent concrete use cases, i.e. models that are foundational to running data systems or that specific communities associated with RDA are considering using as the basis of their data systems.

- **DFT 2: Analysis & Synthesis**  Analyzed the above. 2 major, complementary model categories were noted: ones describing **data organizations** (describing a model) and others focused more on the processing of data according to certain **workflows**. Analytic summaries of each are provided followed by a synthesis which employs a common conceptualization, depicted graphically, to draw a number of conclusions.

- **DFT 3: Term “Snapshot”**  An overview of some core terms & their relations capturing as the DFT WG wrapped up its efforts. Methods and consolidated, core definitions are reviewed based in analysis and synthesis discussed in other documents.
  - The intent of the core snapshot is to used subsequently as a platform to accelerate discussions towards real, working agreements on terminology within RDA and across the worldwide data community.

- **DFT 4: Use Cases**  A collection of use case scenarios developed by the community and discussed at Plenaries as examples of relevant work. These use pertinent term concepts such as PID, Digital Object or Research Data Object. In addition graphics are presented along with additional textual propositions that assert what we should capture in our definitions or issues with concepts.

- **DFT 5: Term Tool Description**

  - This document described the RDA DFT WG Term Definition Tool (aka TeD-T) - a web application for collecting and discussing term definitions. The application is freely available for read access and after a free registration, users are also able to edit existing content or create new entries.
  - The TeD-T platform is deployed and maintained at RZG.
• **Digital Object (aka Digital Entity)** is composed of structured sequence of bits/bytes. As an object it is named. This bit sequence can be identified & accessed by a unique and persistent identifier or by use of referencing attributes describing its properties.
  - Note Digital Entity definition from X.1255 ITU standard “machine-independent data structure consisting of one or more elements in digital form that can be parsed by different information systems; the structure helps to enable interoperability among diverse information systems in the Internet.”

• **Metadata** is a type of data object that that contains attributes describing properties of an associated data or digital object. It may contain as key the persistent identifier of that associated object. The association between a data object and metadata is that the content of the metadata describes the data object. Metadata may serve different purposes, such as helping people to find data of relevance - discovery (Michener 2006) or to bring data together – federation.

• A list of used include:
  - Discovery, Access, Selection, Licensing, authorization, Quality, suitability and Provenance, reproducibility.
  - Data properties, both internal and external, are types of metadata as is transactional information about data.
  - Ref; Michener, W.K. 2006

• **Data Object** is a type of digital object that included the named bits of a digital object but also has representation object allowing processing of its information content.
  - Information that maps a Data Object into more meaningful concepts” (OAIS) — makes humanly-perceptible properties happen
  - Examples: file format, encoding scheme, data format, encoding scheme, data type
• Data Object
  • Type of: Abstract Object (Taxonomy)
  • Sub-types: digital object, …

• Definition: In computer science, an object is any entity that can be manipulated by the commands of a programming language, such as a value, variable, function, or data structure. (With the later introduction of object oriented programming the same word, "object", refers to a particular instance of a class)
  • Definition 2: a Data Object is a dataset

• Equivalent terms (other languages) …

• Attributes….metadata record with data object name, local ID, PID, representation info, checksum…..

• Relations a data element isPartof Data Object…..

• Examples/Instances include: repository metadata, data models, databases, tables, views, files, entities, columns, data elements, and attributes.

• (Source [http://www.indiana.edu/~dss/Services/Naming/nvgglossary.html](http://www.indiana.edu/~dss/Services/Naming/nvgglossary.html))
Coordinated with several other RDA Groups

- **Considerable discussion** of vocabularies has been part of RDA group activities at Plenaries and as part of ongoing RDA group discussion.
- **Cross-group coordinated** with several RDA WGs, as shown in the Data Fabric Figure on data concepts and relations.
  - This coordination ongoing as part of an IG.
  - Potentially all groups could be engaged in this IG and we with them.
- Much more work and discussion would be useful such as with the PP WG and its terminology that was only briefly sketched out without full definitions.
- **PP along with MIG has expressed an interest in more formalized definitions that can be processed by computer and the Ted-T tool may be capable of doing this or at least demonstrating its feasibility.
Ongoing Discussion of the Data LifeCycle define all stages in the existence of digital data from creation to destruction and chained operations Workflows with LC

Core definitions then might include....

- **Curation**: The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use.
  - For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose.
  - Higher levels of Curation will also involve maintaining links with annotation and with other published materials.

- **Archiving**: A curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity.

- **Preservation**: An activity within archiving in which specific items of data/collections are maintained over time so that they can still be accessed and understood through changes in technology.

- **Interoperability**: The ability of a system to accept and send services and to use the services so exchanged to enable them to operate useful. (ISO TC204, document N271)
Notional Core Diagram Reflecting Data Lifecycle and RDA WG

Metadata Types

PID Types

Curation Types?

Provenance Types?

Citation?

Initial Process & Registration

Uses

DTR WG

Uses

Type Registries

Raw Data

Raw Data

Modified by

Modify

Metadata Registries

User/manager

User/manager

PIT WG

PP WG

PIT WG

PP WG

Curation/Provenance policies....

Practical Policies: Divided by Policy Types such as "Manage data sets in a repository"