Data Foundations & Terminology (DFT) WG/IG & VSIG
https://rd-alliance.org/groups/data-foundations-and-terminology-ig.html

Gary Berg-Cross (RDA US Advisory Committee)
RDA P7 Joint VSIG DFT Meeting
Th. March 3rd 9-10:30
Synergy Statement for P7

• There is clear synergy between what the 2 Interest Groups are talking about and with 200+ terms versioned in the DFT term tool (TeD-T) that can serve as a test case for discussing vocabulary services and at the same time advance the consideration of various services in DFT IG.

• To start the DFT vocabulary can provide a number of use cases for discussion in VSIG. These include
  • publish the DFT vocabulary as Linked Data to the Semantic Web (providing URLs to each definition).
  • Another use case is that of vocabulary import - what does it take to export existing DFT vocabulary to a vocabulary server and what parts of vocabularies are easily and what has to be manually edited.
  • A third use case concerns providing more structured relations for the vocabulary.
  • The DFT vocabulary does not include formal taxonomies in the collection, but some services for creating these is available and might be of use.
  • In addition to clarifying discussion some volunteer work to test these ideas and present the results at joint group meetings are possible and under discussion to further advance understanding.
A digital item or group of items referred to as a unit, regardless of type or format that a computer can address or manipulate as a single object.
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)

Data and Digital Objects/Entities

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

Scope
Terms from Model Papers Placed In Tool

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

Analysis and Revision Process

Defs were organized & prepared for review

Active Data

Service Object

Workflow Object

Digital Object

Metadata Object

Representation Object

Information Object

Digital Object of bits and bytes

name
Digital Data Management including unregistered (is a broader concept)

Digital Object Management (registered, digital data)

Where are datasets???
Term definitions have structure in TeD-T.

<table>
<thead>
<tr>
<th>Access</th>
<th>Access control list</th>
<th>Add a retention period</th>
<th>Aggregation</th>
<th>Authentication</th>
<th>Authorize a deposition</th>
<th>Blueprint</th>
<th>Cataloguing</th>
<th>Citable Data</th>
<th>Collection Management Identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access Workflow</td>
<td>Active Collection</td>
<td>Addition of access controls</td>
<td>Architecture</td>
<td>Authenticity metadata</td>
<td>Bit Sequence</td>
<td>Canonical Data Collection</td>
<td>Checksum</td>
<td>Citation Metadata</td>
<td>Components</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Access a repository</td>
<td>Active Data</td>
<td>Administrative metadata</td>
<td>Attribute</td>
<td>Authoritative source</td>
<td>Bit Stream</td>
<td>Catalog</td>
<td>Choosing a storage location</td>
<td>Content Interpretation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Definition:**

**Explanation:**

**Examples:**

**References:**

**Scope:**

**Status:**
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>
Policy Components - Conceptual Fundamentals

Policy-based Data Management Concept Graph

Purpose
  - SubType
  - Sharing
    - Publication
    - Preservation
      - Integrity
    - Authentication
    - Access control
      - HasFeature
      - Completeness
      - Correctness
      - Consensus
      - Consistency
  - Defines
      - Replication Policy
      - Checksum Policy
      - Quota Policy
      - Data Type Policy

Collection
  - Has
    - Digital Object
      - Isa
      - Has
      - Updates
      - Isa

Property
  - Isa
  - HasFeature
  - Defines
    - Policy
      - Isa
      - Has
      - Control
      - SubType
        - HasFeature
        - Periodic Assessment Criteria Policy
          - Isa
          - Invokes
            - Policy Enforcement Point
              - Isa

Procedure
  - Isa
  - Has
    - Workflow
      - Isa
      - Chains
      - Isa

Function
  - Isa
  - Operation

Attribute
  - Isa
  - Updates

Persistent State Information
  - Isa
  - Isa
  - Isa
  - Isa
  - Isa
  - Isa
  - Isa
  - Isa
  - Isa
  - Isa
  - Isa
  - Isa
  - Isa
  - Isa
  - Isa

research data sharing without barriers
rd-alliance.org
Background & Match up with VSIG Objectives

DFT is building an RDA data vocabulary, but leveraging others efforts too.

• TeD-T Term Definition Tool:
  http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page

1. We are cooperating with VSIG on its survey vocabulary efforts from related communities (Provenance IG, Research Administration Information (CASRAI) interactive Glossary, ISO 5127 Standard Information & documentation -- Vocabulary --): Acquisition, identification, and analysis of documents and data etc.)

2. We want to publish our vocabulary for more people to use

3. We are interested in identifying common functionality for vocabulary publication services

4. We have understood some functions for Voc service that would serve us and they are in our Use Cases

5. We have started on a set of 10 uses for a Voc service

6. As a test case DFT could help VSIG develop recommendations for vocabulary publication services
What Problem(s) will Voc Services help DFT with?

• Our Uses cases help identify the problems that we think that DFT need help with improve the quality of the DFT vocabularies
  • Add synonyms, URIs for each term, handle taxonomy, etc....

1. Exporting existing DFT vocabulary to a server like RVA. Exercises APIs but requires formatting in SKOS.

2. Creating one or more DFT taxonomies from the DFT vocabulary collection.
   The following are sub-use cases as part of managing the Thesauri:
   - Creating Relations by Drag and Drop and move Concepts by Drag and Drop
   - Merging Concepts by Drag and Drop
   - Adding Relations Using Autocomplete
   - Adding Notes to your Concepts (or import notes from DFT tool)

3. Upload relevant sub-sample of DFT documents used for vocabulary development and enhance the existing collection by analysis of a relevant sub-sample of DFT documents and the RVA products:
   - Candidate Terms List, Extracted Concepts List, Extracted Terms List.

4. Test use of the Custom Scheme to creation custom classes, relations and attributes:

5. Use defined relation types in Voc Servier relating concepts.

6. Adding attributes to defined concepts using VS.

7. Exploring the use of any predefined Ontologies in a VS to enhance the DFT vocabulary.

8. Create a custom ontology from a portion of the DFT vocabulary using VS.

9. Publish the DFT vocabulary as Linked Data to the Semantic Web. (proving URLs to each definition)
Use cases are posted as Google docs:
https://rd-alliance.org/group/vocabulary-services-interest-group/wiki/community-use-cases.html

- Export existing RDA DFT vocabulary to RVA
- Create a concept collection
- Get definition, source and labels for a concept given the URI
- Select ConceptURI to identify term
- Get a list of all transitive relations used in the register....
Use Case: Export existing RDA DFT vocabulary to RVA

Point of Contact: Gary Berg-Cross <gbergcross@gmail.com>

Version: V.1

Date: 1/15/16

Use Case Name
Export DFT vocabulary

Goal
Export existing RDA DFT vocabulary to RVA

Summary
Exporting existing DFT vocabulary to RVA is the first step to test the value of the RVA for DFT. It will exercise the 2 APIs. The DFT tool is built on the Semantic Media Wiki and can export in an RDF form. What is interesting here is to see what information from the DFT tool can be imported properly and what has to be cut and pasted etc., to make it usable for other things such as taxonomy building.
Following Jane’s presentation looked at the tool to start on an import
• Did an RDF export as step 1
• Looking at SKOS requirements to make file acceptable

skos:prefLabel "Root vegetable"@en;
dcterms:created "2015-05-14T02:07:30Z"^^xsd:dateTime ;
dcterms:creator "janeAdmin" ;
dcterms:modified "2015-05-14T02:07:30Z"^^xsd:dateTime ;
skos:altLabel "tuber"@en ;
skos:definition "Root vegetables are underground plant parts used as vegetables."@en ;
Handling upper level documentation & details in the definitions

RDA info (Thomas, datetime etc>)
including a Subject since we are dealing with data vocabularies and not vegetables:

dcterms:description "A vocabulary of vegetables."@en ;
dcterms:contributor "janeAdmin" ;
dcterms:publisher "World Vegetable Organisation" ;
dcterms:subject "Food"@en .

A Term like "digital object" has definition(s) and a label but not something yet like a narrower term so we ignore these for now.

**SKOS Profile Idea**
And we need to include a link to Explanation of definition and Example of definition. But it looks like the SKOS Note idea might serve for this.