RESEARCH DATA ALLIANCE

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

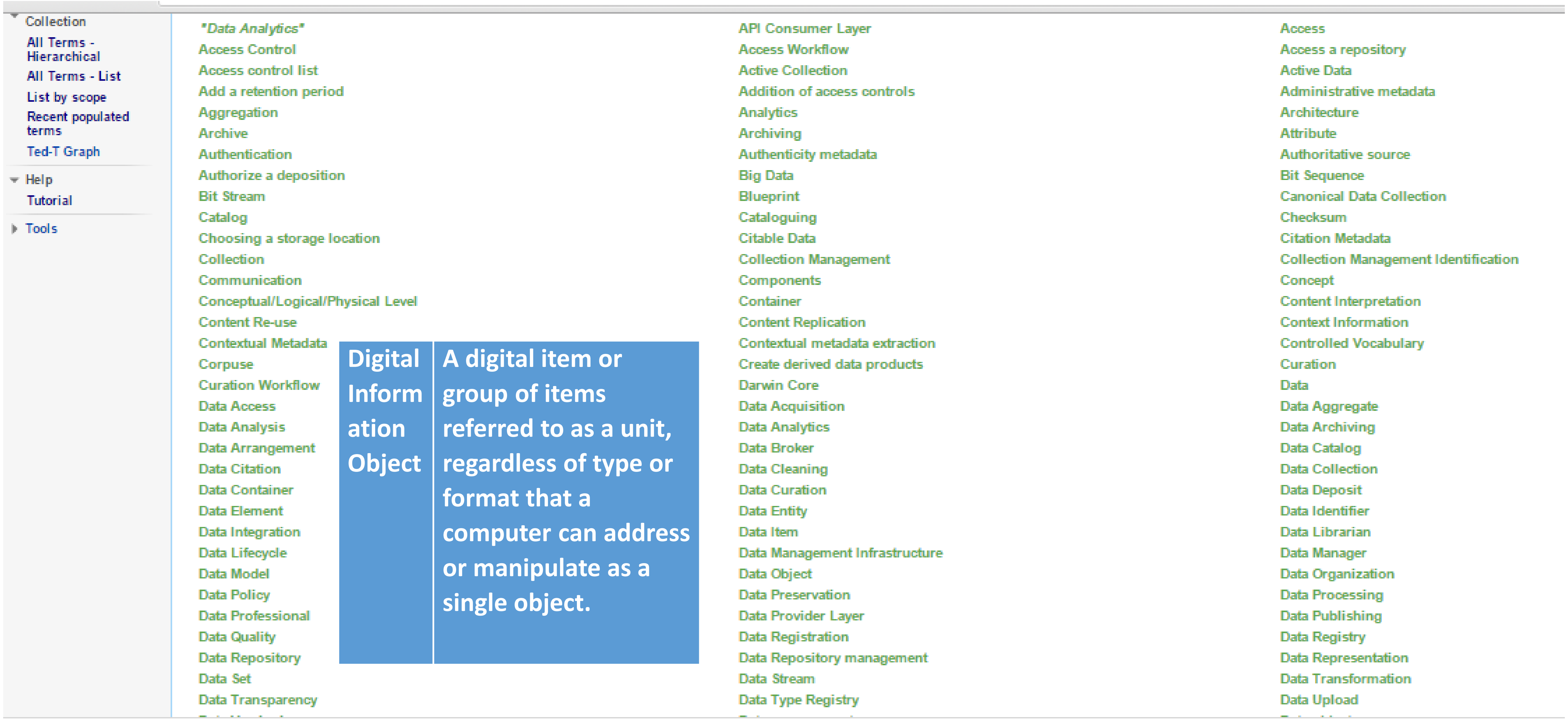
research data sharing without barriers
rd-alliance.org

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.

Portion of Terms in TeD-T

http://smw-rda.esc.rzg.mpg.de/index.php/Special:AllPages



"Data Analytics"
Access Control
Access control list
Add a retention period
Aggregation
Archive
Authentication
Authorize a deposition
Bit Stream
Catalog
Choosing a storage location
Collection
Communication
Conceptual/Logical/Physical Level
Content Re-use
Contextual Metadata
Corpuse
Curation Workflow
Data Access
Data Analysis
Data Arrangement
Data Citation
Data Container
Data Element
Data Integration
Data Lifecycle
Data Model
Data Policy
Data Professional
Data Quality
Data Repository
Data Set
Data Transparency

API Consumer Layer
Access Workflow
Active Collection
Addition of access controls
Analytics
Archiving
Authenticity metadata
Big Data
Blueprint
Cataloguing
Citable Data
Collection Management
Components
Container
Content Replication
Contextual metadata extraction
Create derived data products
Darwin Core
Data Acquisition
Data Analytics
Data Broker
Data Cleaning
Data Curation
Data Entity
Data Item
Data Management Infrastructure
Data Object
Data Preservation
Data Provider Layer
Data Registration
Data Repository management
Data Stream
Data Type Registry

Access
Access a repository
Active Data
Administrative metadata
Architecture
Attribute
Authoritative source
Bit Sequence
Canonical Data Collection
Checksum
Citation Metadata
Collection Management Identification
Concept
Content Interpretation
Context Information
Controlled Vocabulary
Curation
Data
Data Aggregate
Data Archiving
Data Catalog
Data Collection
Data Deposit
Data Identifier
Data Librarian
Data Manager
Data Organization
Data Processing
Data Publishing
Data Registry
Data Representation
Data Transformation
Data Upload

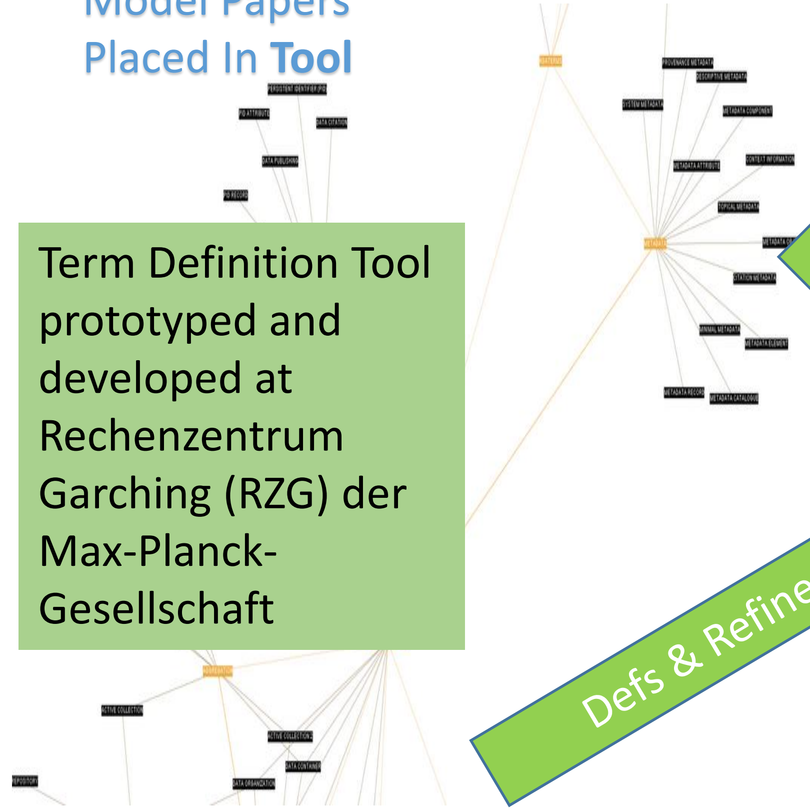
Digital Information Object	A digital item or group of items referred to as a unit, regardless of type or format that a computer can address or manipulate as a single object.

Scope
Terms from
Model Papers
Placed In **Tool**

Overview of Term Development

Defs were organized & prepared for review

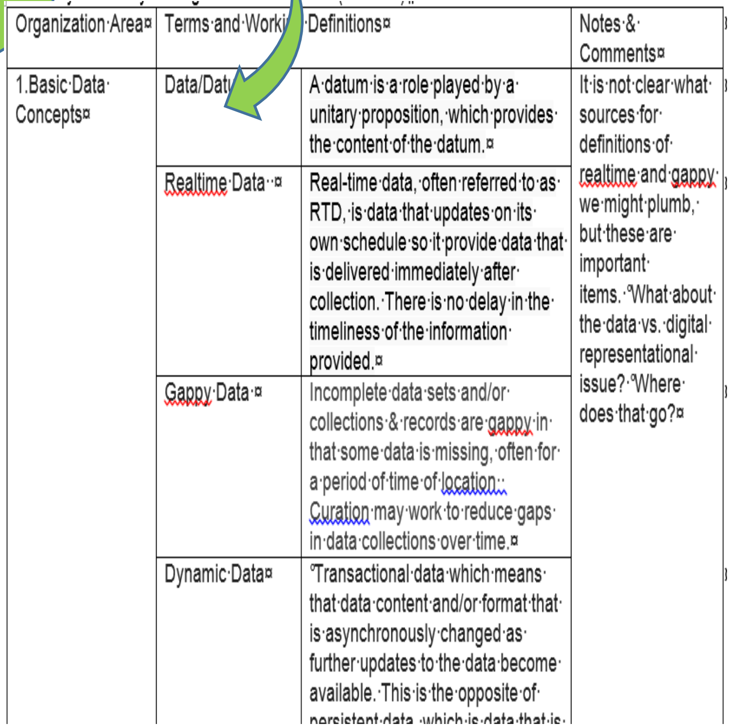
Analysis and Revision Process



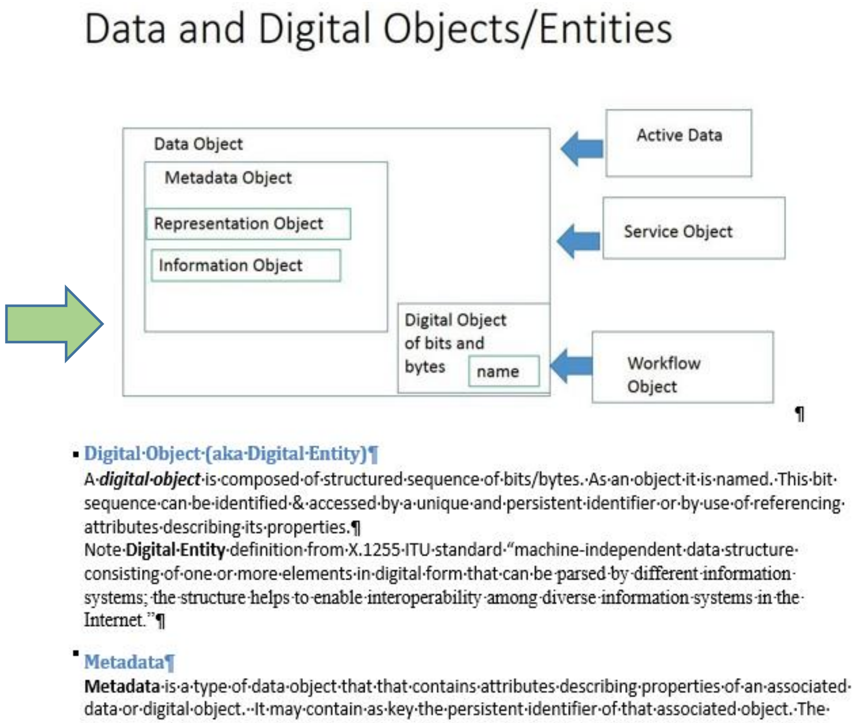
Term Definition Tool prototyped and developed at Rechenzentrum Garching (RZG) der Max-Planck-Gesellschaft

Defs & Refinement

Digital Object	A *digital object* 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.



Organization·Area¤	Terms·and·Worki	Definitions¤	Notes·&·Comments¤
1.Basic·Data·Concepts¤	Data/Datu	A·datum·is·a·role·played·by·a·unitary·proposition,·which·provides·the·content·of·the·datum.¤	It·is·not·clear·what·sources·for·definitions·of·realtime·and·gappy·we·might·plumb,·but·these·are·important·items.·°What·about·the·data·vs.·digital·representational·issue?·°Where·does·that·go?¤
	Realtime·Data··¤	Real-time·data,·often·referred·to·as·RTD,·is·data·that·updates·on·its·own·schedule·so·it·provide·data·that·is·delivered·immediately·after·collection.·There·is·no·delay·in·the·timeliness·of·the·information·provided.¤	
	Gappy·Data·¤	Incomplete·data·sets·and/or·collections·&·records·are·gappy·in·that·some·data·is·missing,·often·for·a·period·of·time·of·location.:·Curation·may·work·to·reduce·gaps·in·data·collections·over·time.¤	
	Dynamic·Data¤	°Transactional·data·which·means·that·data·content·and/or·format·that·is·asynchronously·changed·as·further·updates·to·the·data·become·available.·This·is·the·opposite·of·persistent·data,·which·is·data·that·is·	



Data and Digital Objects/Entities

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

¶

- **Digital·Object·(aka·Digital·Entity)¶**

  A·***digital·object***·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¶**

  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·

Starter areas and items :

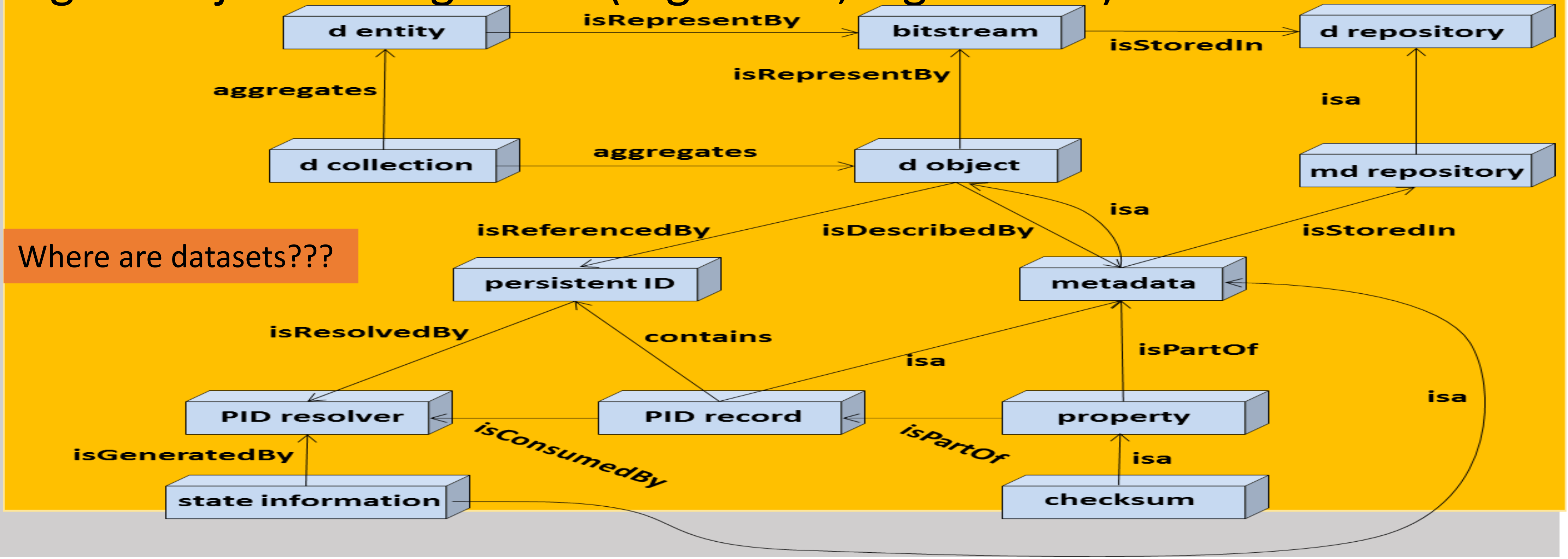
Persistent Identifiers (PIDs and types)
Digital Object - Data Object
Collection - Data Set - Aggregation

Repository (Registries and related **Policies**)

Concept map overview of Core Terms
Broadening the Discussion (Stepwise or Scope-wise)

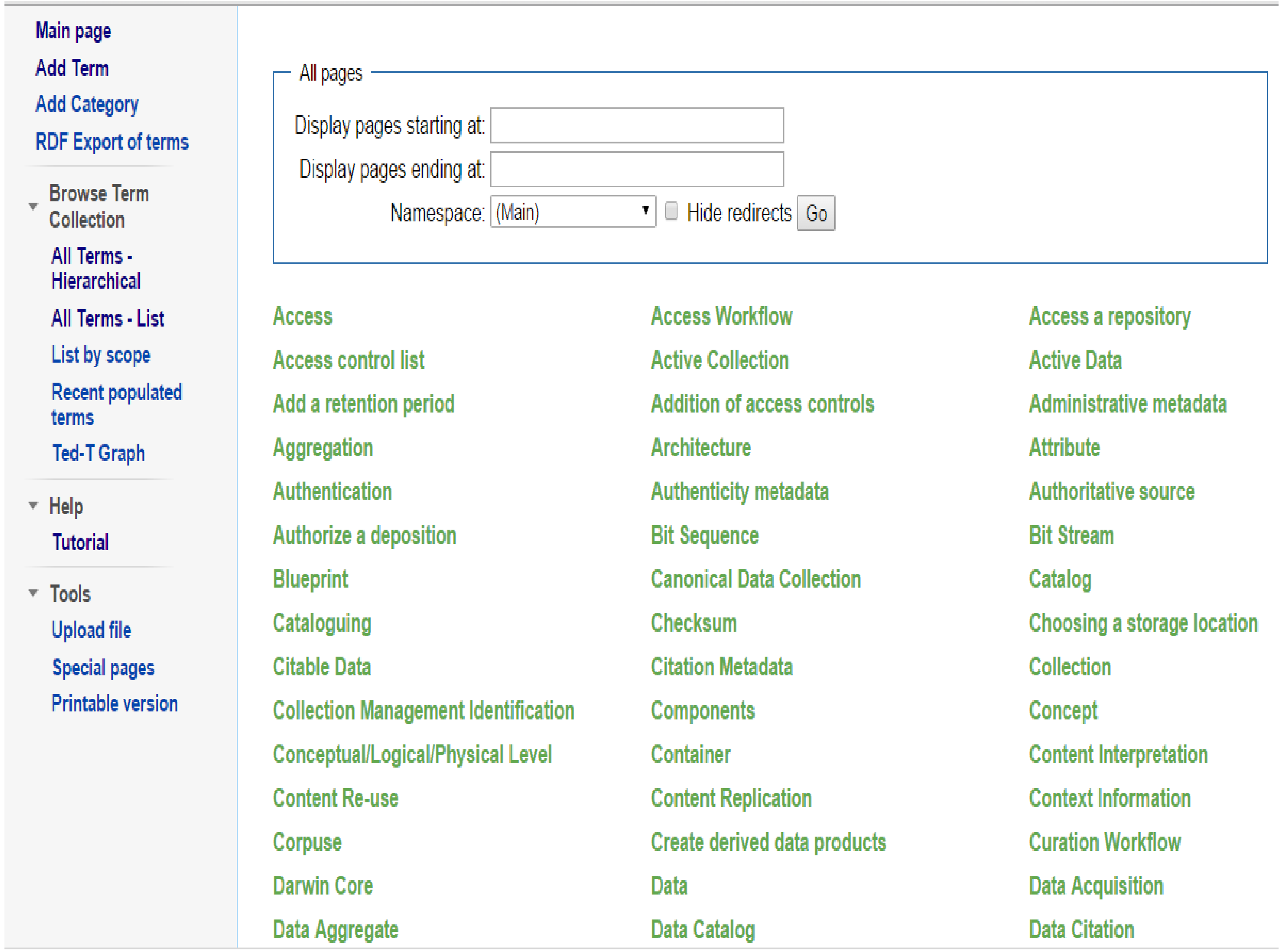
Digital Data Management including unregistrered (is a broader concept)

Digital Object Management (registered, digital data)



Where are datasets???

Term definitions have structrure in TeD-T



- Main page
- Add Term
- Add Category
- RDF Export of terms

Browse Term Collection

- All Terms - Hierarchical
- All Terms - List
- List by scope
- Recent populated terms
- Ted-T Graph

Help

- Tutorial

Tools

- Upload file
- Special pages
- Printable version

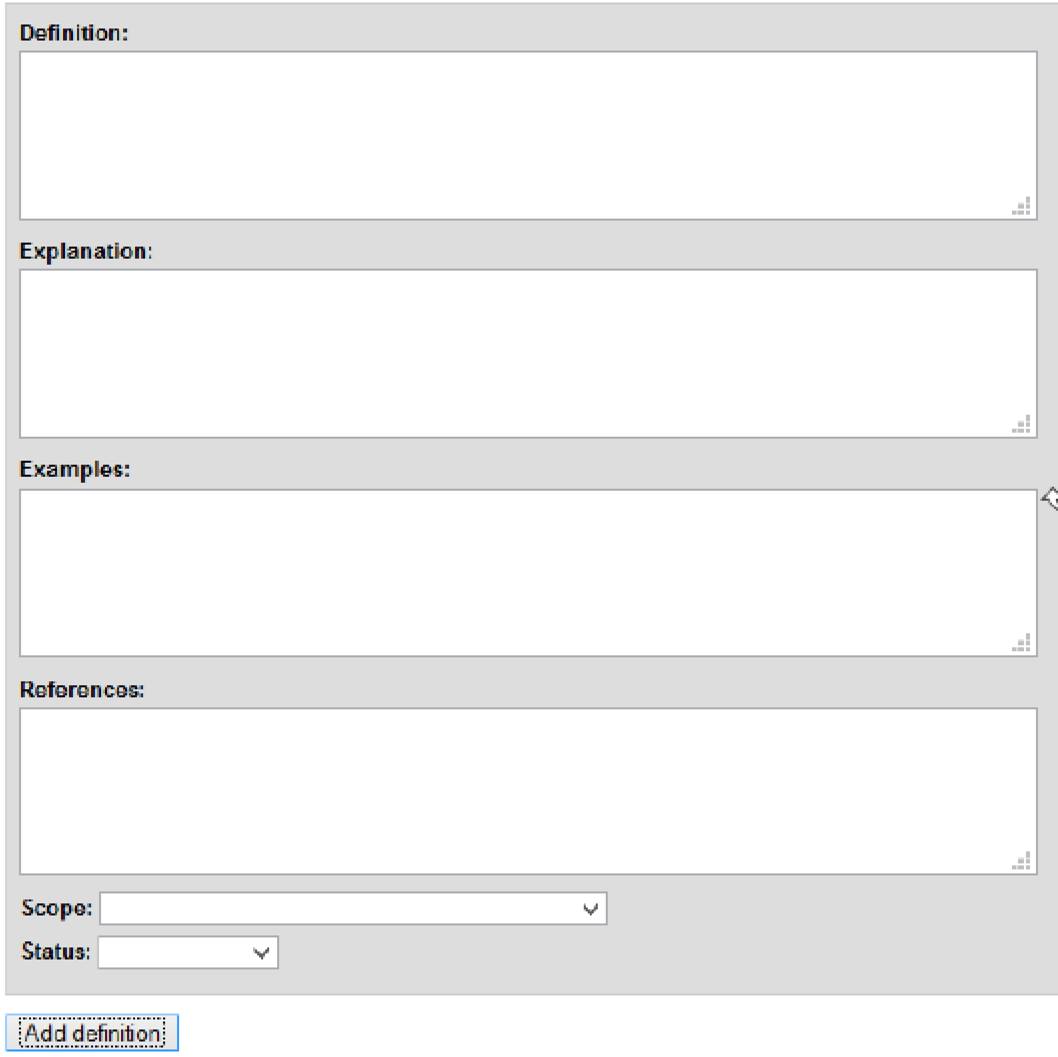
All pages

Display pages starting at:

Display pages ending at:

Namespace: (Main) ▾ □ Hide redirects Go

- Access
- Access Workflow
- Access a repository
- Access control list
- Active Collection
- Active Data
- Add a retention period
- Addition of access controls
- Administrative metadata
- Aggregation
- Architecture
- Attribute
- Authentication
- Authenticity metadata
- Authoritative source
- Authorize a deposition
- Bit Sequence
- Bit Stream
- Blueprint
- Canonical Data Collection
- Catalog
- Cataloguing
- Checksum
- Choosing a storage location
- Citable Data
- Citation Metadata
- Collection
- Collection Management Identification
- Components
- Concept
- Conceptual/Logical/Physical Level
- Container
- Content Interpretation
- Content Re-use
- Content Replication
- Context Information
- Corpuse
- Create derived data products
- Curation Workflow
- Darwin Core
- Data
- Data Acquisition
- Data Aggregate
- Data Catalog
- Data Citation



Definition:

Explanation:

Examples:

References:

Scope:

Status:

Add definition

Practical Policy WG area examples

- Contextual metadata extraction

- Data access control
- Data backup
- Data format control
- Data retention
- Disposition
- Integrity (including replication)
- Notification..

- A start on minimal MD?
- Key processes across the data lifecycle?

Extract metadata	Attribute_name
	Attribute_value
	Attribute_unit
	Source_file
	Source_collection

Contextual metadata extraction policies

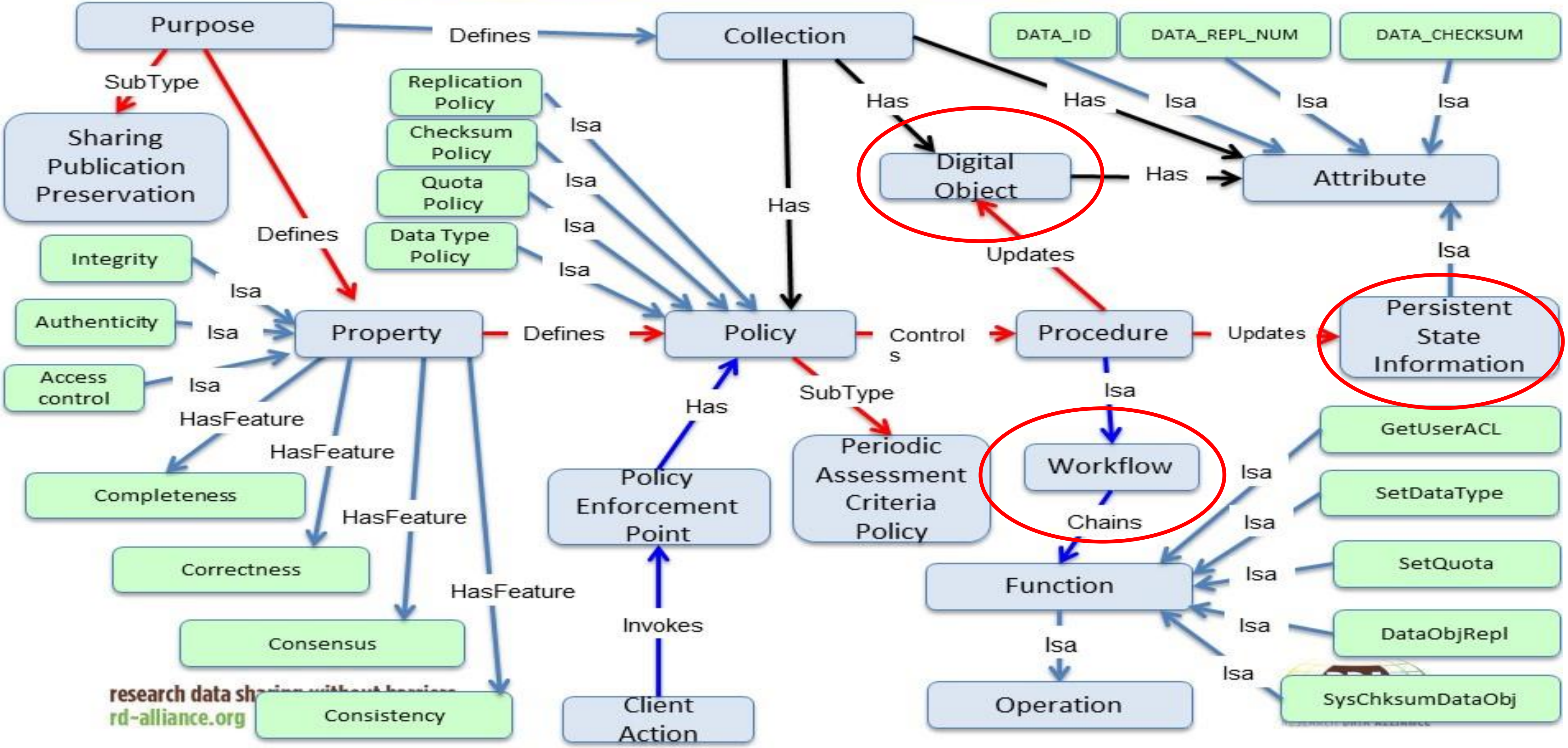
This policy area focuses on metadata associated with files and collections.

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.

Policy Components - Conceptual Fundamentals

Policy-based Data Management Concept Graph


Purpose
Defines
Collection
DATA_ID
DATA_REPL_NUM
DATA_CHECKSUM
SubType
Sharing Publication Preservation
Replication Policy
Checksum Policy
Quota Policy
Data Type Policy
Isa
Isa
Isa
Isa
Has
Has
Isa
Isa
Isa
Digital Object
Has
Attribute
Has
Defines
Integrity
Authenticity
Access control
Isa
Isa
Isa
Updates
Isa
Property
Defines
Policy
Controls
Procedure
Updates
Persistent State Information
HasFeature
HasFeature
HasFeature
HasFeature
Has
SubType
Isa
Completeness
Correctness
Consensus
Consistency
Policy Enforcement Point
Invokes
Client Action
Periodic Assessment Criteria Policy
Workflow
Chains
Function
Isa
Operation
GetUserACL
SetDataType
SetQuota
DataObjRepl
SysChksumDataObj
Isa
Isa
Isa
Isa
Isa

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 identifing common functionality for vocabulary publication services
4. We have understand 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?

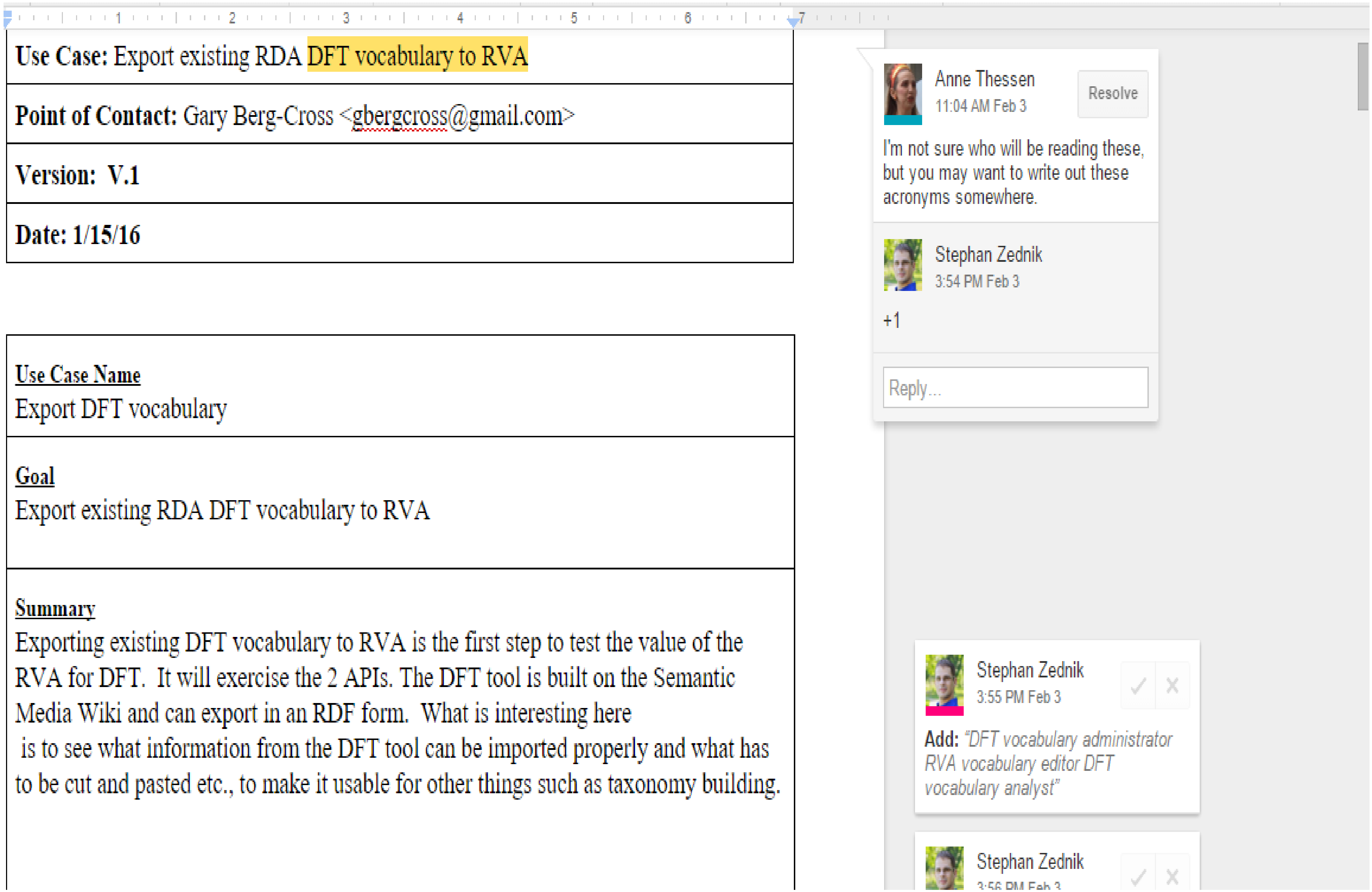
- Our Uses cases help identify the problems that we think that DFT need help with improve the quality of the DFT vocabularies
  - Add synonyms, URIs for each term, handle taxonomy, etc....

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.

Anne Thessen
11:04 AM Feb 3

I'm not sure who will be reading these, but you may want to write out these acronyms somewhere.

Stephan Zednik
3:54 PM Feb 3

+1

Stephan Zednik
3:55 PM Feb 3

Add: *"DFT vocabulary administrator RVA vocabulary editor DFT vocabulary analyst"*

Stephan Zednik
3:56 PM Feb 3

Work with Research Vocabulary Australia (RVA)

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

```
<https://editor.vocabs.ands.org.au/examplepoolpartyproject/88> a skos:Concept ;
skos:prefLabel "Root vegetable"@en ;
skos:topConceptOf <https://editor.vocabs.ands.org.au/examplepoolpartyproject/87> ;
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 ;
skos:narrower <https://editor.vocabs.ands.org.au/examplepoolpartyproject/89> .
```

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 vegtables:

```
<https://editor.vocabs.ands.org.au/examplepoolpartyproject/112> ;
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.