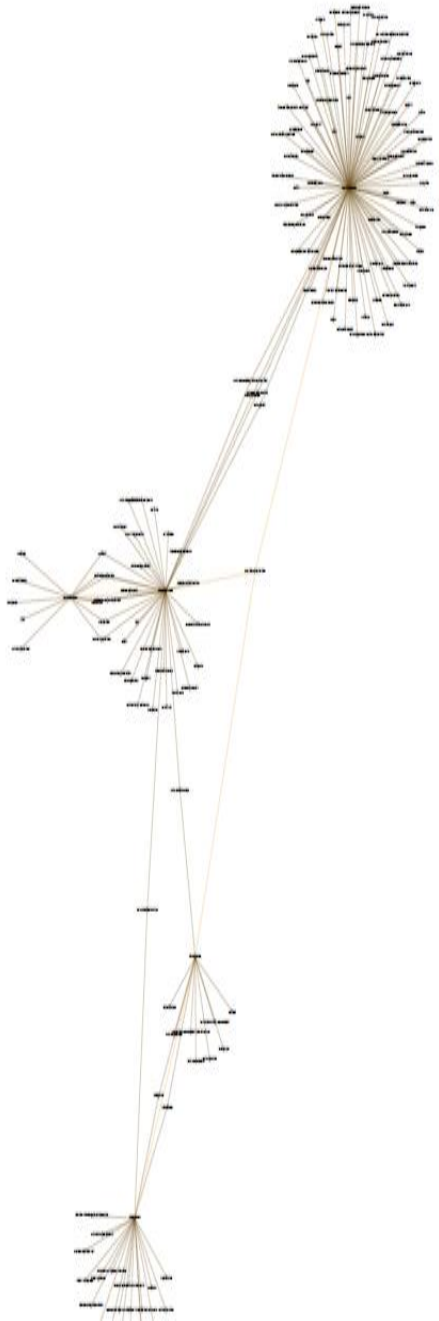
RESEARCH DATA ALLIANCE

RDA Data Foundation and Terminology (DFT) IG: Introduction

research data sharing without barriers
rd-alliance.org



- Prepared for RDA 7th Plenary Tokyo, March 2, 2016
- Gary Berg-Cross

DFT IG Session Wed.(11:00-12:30) Agenda

- Overview of the DFT IG, Case Statement & the Breakout Session- Goals and Plans Gary Berg-Cross
  - We are now an officially recognized group
- Overview of the Ted-T tool Status (Raphael Ritz)
- Discussion of Liaison relation to other RDA IGs and WGs
  - Data Fabric
  - Metadata (MD profile)
  - Vocabulary Services
  - Provenance
  - Data Publishing Workflow
  - Practical policy
  - Others – water vocabulary

- General Discussion
- Discussion of follow on work & Plan for follow up virtual meetings.

What Problem(s) are we trying to help with?

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.

- Among the problems
  - RDA efforts to make data sharing easier
  - Data organizations (DOrg) and ideas about it are all different
  - We are all using different vocabularies, wasting time and misunderstanding each other in a variety of data projects
  - Different DOrgs make data discovery and integration very time consuming, inefficient and thus expensive
  - Different DOrgs prevent us developing maintainable data infrastructure & support software
- There is a wide impacted across domains, professions, etc.
  - All efforts to integrate and make data open
- What are the ramifications of not having the problem resolved?
  - Combining data of all sorts across different origins (projects, repositories, disciplines, etc.) is a nightmare and requires a lot of curation and transformation before the actual scientific analysis can start
  - Interoperability remains a challenging goal.

More Information on Products

- DFT WG:
https://rd-alliance.org/groups/data-foundation-and-terminology-wg.html
- DFT IG:
https://rd-alliance.org/groups/data-foundations-and-terminology-ig.html
- DFT IG Case Statement:
https://www.rd-alliance.org/group/data-foundations-and-terminology-ig/case-statement/case-statement.html
- TeD-T Term Definition Tool:
http://smw-rda.esc.rzg.mpg.de/index.php/Main_Page

Case Statement, TAB & Groups

- We have addressed TAB issues in an updated version of the Case Statement
- Added to our statement the discussions with people from international efforts like
  - Data Publication Workflow group and members, such as Walter Stewart, of Research Data Canada
  - Science Europe Working Group on Research Data (Peter Doorn) and
- Paul Millar's work funded under H202 project INDIGO-DataCloud to define the terms a user-community uses when describing the expected Quality-of-Service and Data-Lifecycle of their storage infrastructure.

  They held a QoS/DataLS session Tuesday.
- Doing work with Vocabulary Services.

Terminology Issue(s) : What do we expect from RDA ?

Adopt one or build own language?

- Build our own language stepwise,
  - Use a tool to help
- And cooperate with RDA Groups.....other efforts

Effective community discussion to avoid spending interminable time on terminology debates that can‘t close

Portion of Terms in TeD-T

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

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

Help
- Tutorial

Tools

"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

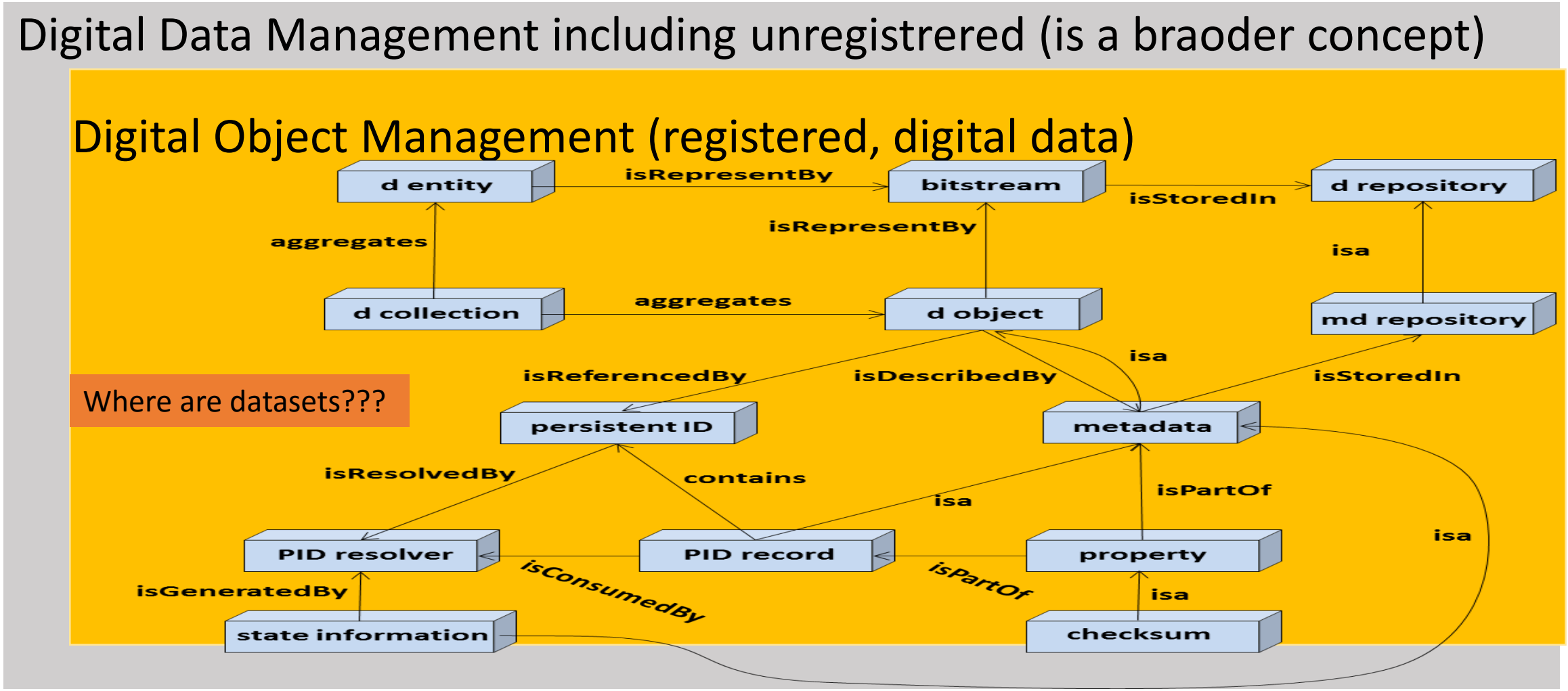
Digital Information Object	**A digital item or group of items referred to as a unit, regardless of type or format that a computer can address or manipulate as a single object.**

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

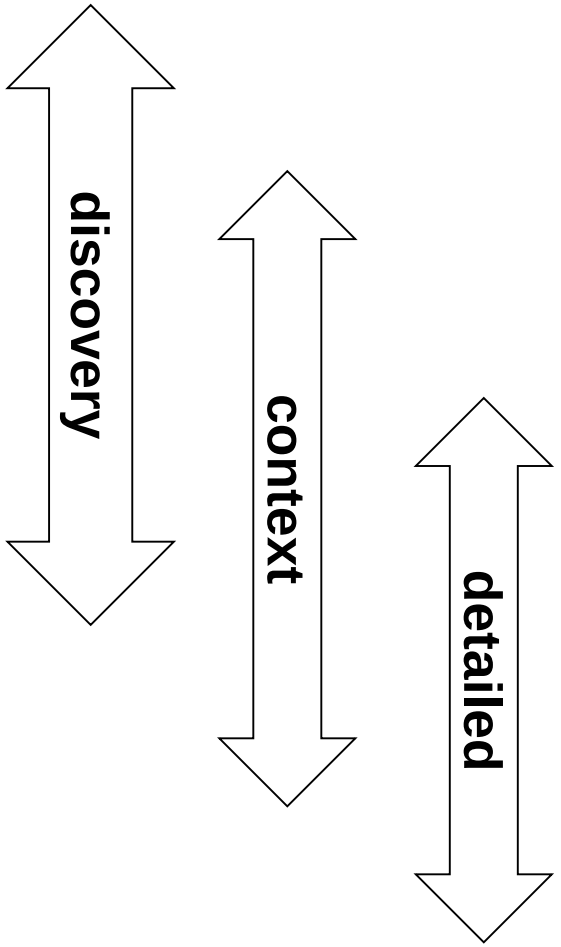
Concept map overview of Core Terms Broadening the Discussion (Stepwise or Scope-wise)

Data Management (and use) is broader still



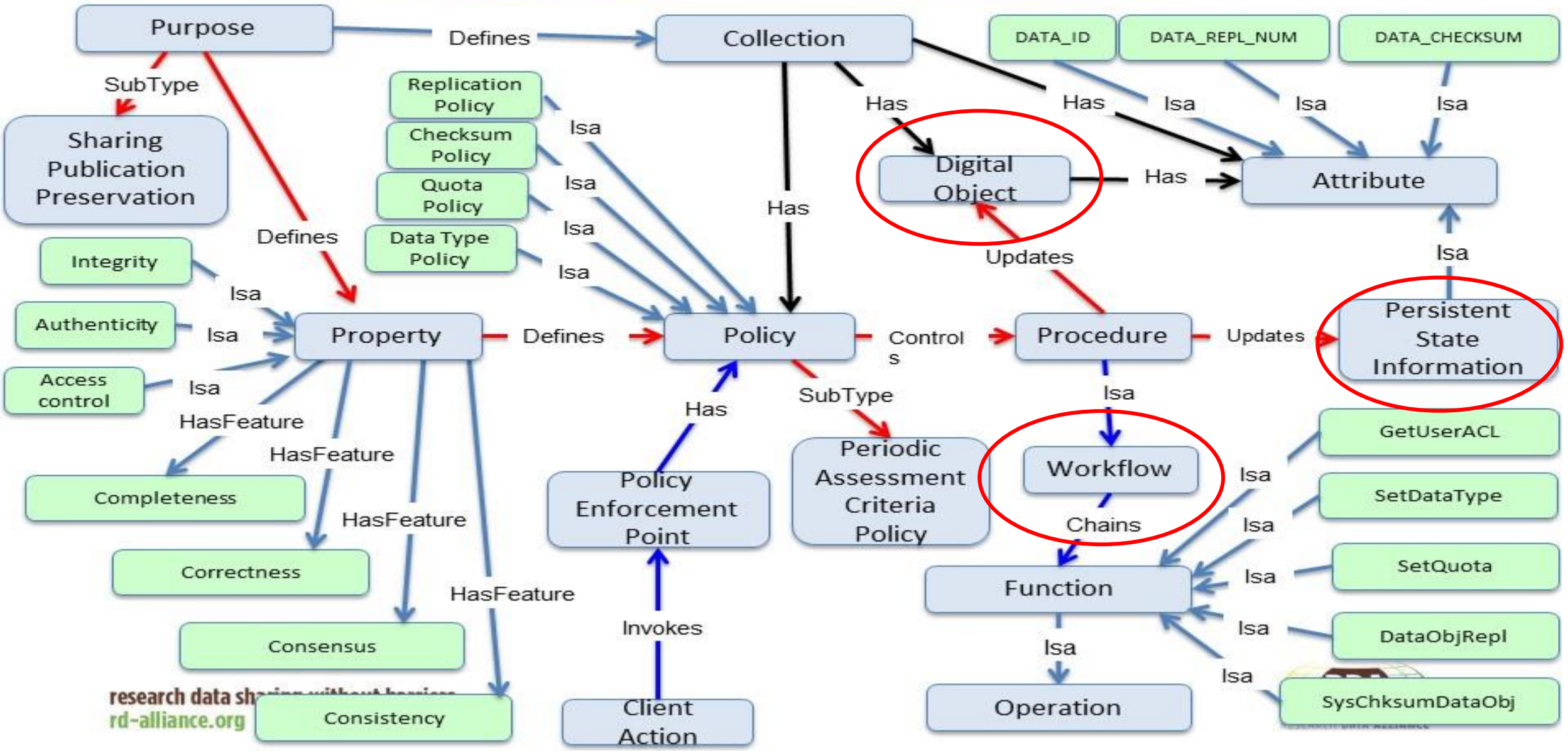
Input from MD Groups:
Open Data: Purposing the elements

- Unique Identifier (for later use including citation)
- Location (URL)
- Description
- Keywords (terms)
- Temporal coordinates
- Geospatial coordinates
- Originator (organisation(s) / person(s))
- Project
- Facility / equipment
- Quality
- Availability (licence, persistence)
- Provenance
- Citations
- Related publications (white or grey)
- Related software
- Schema
- Medium / format



Policy Components - Conceptual Fundamentals

Policy-based Data Management Concept Graph



Responded to Data Fabric concepts with candidate terminology: Examples

1. Data practice is the actual application/ use of ideas & methods (as opposed to theories) about how data are collected, created, stored (maintained), curated, used, shared and released (disseminated).
2. Data principles are rules that provide guidance across data management and use for such things as" data acquisition, data lifecycle control, data policy & ownership, metadata practices, data quality etc.
3. Common data solutions are agreed upon, easily available, tested & approved approaches to widely occurring problems in data management and use
4. Data discovery is a process of query and/or search to find (research) data of interest.
5. Database cracking features incremental partial indexing and/or sorting of the data. It combines features of automatic index selection and partial indexes.
   - It reorganizes data within the query operators, integrating the re-organization effort (occasionally invoking creation or removal of indexes on tables and views based on use) into query execution.
   - It shifts the cost of index maintenance from updates to query processing.
6. Adaptive indexing is characterized by the partial creation and refinement of preliminary or fixed DB indexes as side effects to support efficient query execution. (after http://www.vldb.org/pvldb/vol4/p586-idreos.pdf)

Adding New Terms/Defs from Data Fabric

Group	Term	Comment - terms from DICE group	Reference
	Accessible	Can be retrieved through the Internet	
	Actor ID	Unique identifier for an operation	
	Actors	Operations applied to digital objects	
	Aggregation	Physical placement in a container, or logical organization of name space	DFT
	Architecture		
	Architecture Framework		ISO/IEC/IEEE 42010
	Architecture View		ISO/IEC/IEEE 42010
	Architecture Viewpoint		ISO/IEC/IEEE 42010
	Authentication System	Verification of identity	
	Authorization System	Specification of permission to do desired operation	
	Bitstream	Sequence of bits	DFT
	Catalogue	System for managing names and related attributes	
	Category Registry	System for registering catalogues	
	Certification	Assertion that a set of properties are correctly enforced	
	Checksum	Unique reduced representation of a digital object	DFT
	Citation	Reference to a uniquely identified digital object	
	Collection Registry	Register of collections	
	Common Component	Software implementing an operation that is used in multiple systems	
	Common Policies	Assertions about management of a system used by multiple institutions	
	Component	Software implementing an operation	
	Configuration	Organization of components into a specific arrangement	

Input from Reagan Moore

DFT as a Test Case for Vocabulary Services IG

VSIG Objectives

1. VSIG will **survey** vocabulary efforts from related communities and provide a summary report on existing practices
    1. DFT is a vocabulary effort with > 200 terms
2. VSIG will identify common terminology in the context **of vocabulary publication**
    1. We want to publish our vocabulary for more people to use
3. VSIG will identify **common functionality** for vocabulary publication services
    1. We have understand some functions for Voc service that would serve us and they are in our Use Cases
4. VSIG will collect and report on **use cases** from the research community regarding using published vocabularies and vocabulary services
    1. We have started on a set of 10 uses for a Voc service
5. VSIG will develop recommendations for vocabulary publication services
    - DFT would assist in that

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.

Use Cases

https://rd-alliance.org/group/vocabulary-services-interest-group/wiki/community-use-cases.html

- 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.
  - In some it is more than converting to RDF. There is a SKOS profile needed by some
- Publish the DFT vocabulary as Linked Data to the Semantic Web
- Providing URLs to each DFT definition.
- Providing more structured relations for the vocabulary.
- Adding a **formal taxonomies** in the collection
- Merging 2 vocabularies, e.g. DFT and Provenance terms

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.

Objectives for P8

1. Continue IG discussion and leverage existing work and approach but improve both
    1. We are expecting considerable discussion of new requirements coming out of groups nearing completion, but also support as part of adoption.
    2. We hope to leverage Vocabulary Services to improve the quality of the work
    3. We can also leverage the experience of other IGs as to success factors
2. Focus on facilitating community discussion on core concepts
    1. Based on feedback, some curated revisions on definitions and extension of the current synthesis model can be expected to finalize and stabilize the effort for subsequent use.
3. Facilitate approach for definition development
    1. Potential adopters will be encouraged at P7 to provide feedback on additional use case scenarios to illustrate what areas of work they plan on using the models and vocabulary for.
    2. This will serve to plan work and virtual meetings between P8 and P9.