RESEARCH DATA ALLIANCE

Working Group:
Data Foundation and Terminology (DFT)

Peter Wittenburg, Gary Gerg-Cross, Raphael Ritz

research data sharing without barriers
rd-alliance.org

Purpose & Use - DFT Work Group

The DFT WG task is to:

- 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.

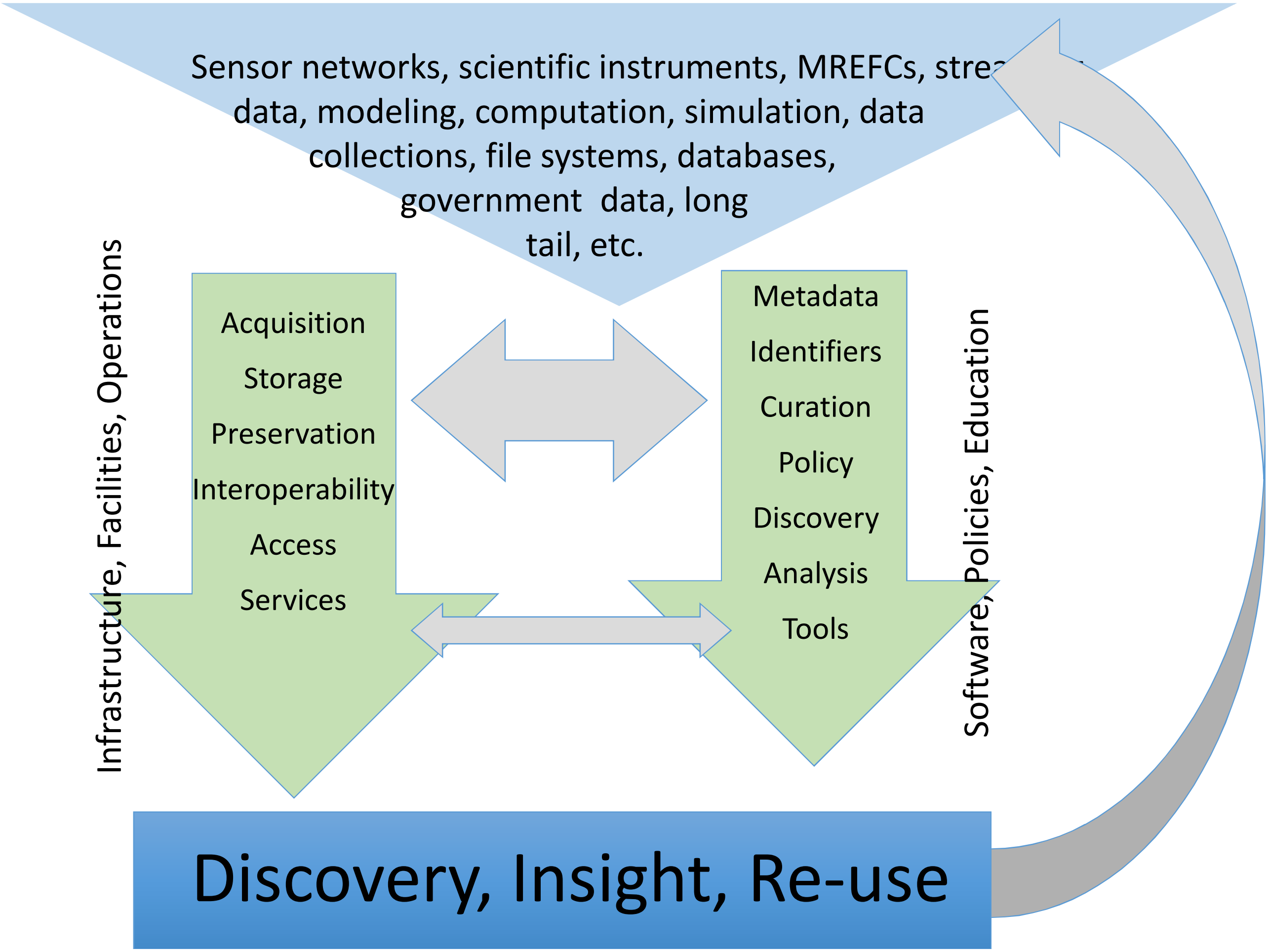
The model and its derived reference data should be sound, practical and agreed to within the community for use:

- across communities and stakeholders to better synchronize data conceptualization,
- to enable better understanding within and between communities, and
- to stimulate tool building, such as for data services, supportive of the basic model's use.
  - Need to get the story straight on model to govern the use of related tools.

Metadata
Identifiers
Curation
Policy
Discovery
Analysis
Tools

Data access – Requests for data services, such as a query of a Data Asset. These requests are supported by Data Access Services (**Data.Gov Data Reference Model)**

We are Talking about Data Management



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

Infrastructure, Facilities, Operations

Acquisition

Storage

Preservation

Interoperability

Access

Services

Metadata

Identifiers

Curation

Policy

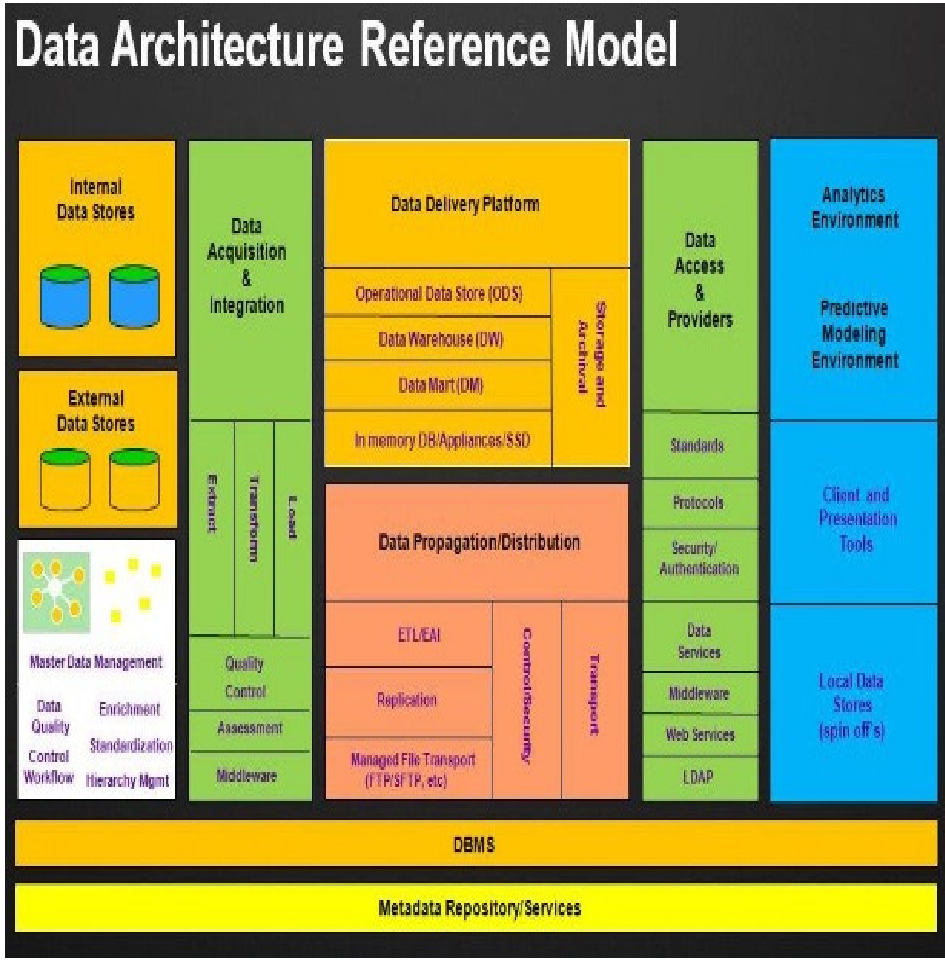
Discovery

Analysis

Tools

Software, Policies, Education

Discovery, Insight, Re-use



Affordance Opportunity -Analogous to Internet Protocol

The IP reference model established a standard language of networking, which in turn:

- sharpened the global understanding of the need for systematic relations of the various protocol layers and
- enabled basic protocol notions to be realized by such things as IP and TCP.

In the new era of DATA we need something comparable for that Domain.

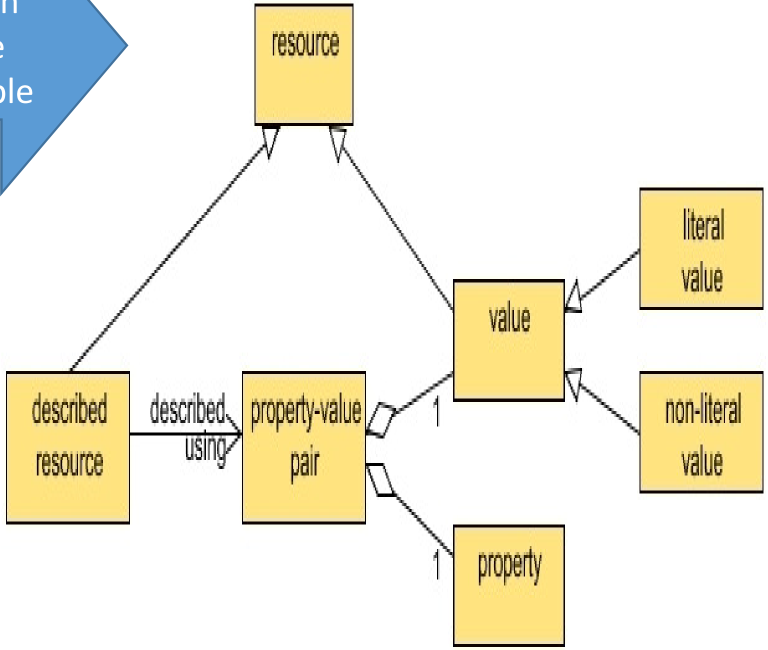
Example problem due to lack of Shared Vocabularies (from Mark Alan Parsons)

- NSIDC and several groups at NCAR began a collaborative project ~5 years ago.
- NSIDC typically considers a "data set" as a collection of related files.
- NCAR was thinking in the THREDDS context where a data set is essentially a file.
- After working together for months and struggling to understand each other about a variety of topics the groups kept talking past each other until
  - in one meeting they had an aha moment &realized they were using the term "data set" in very different ways.
  - The core of the issue was **granularity**. THREDDS is very granular. NCAR were more aggregated. This has a big affect on how you describe, share, manage etc. the "data set"
- After that, things got a lot clearer and collaboration improved.

WG Tasks - be finished in about 15 -18 months

1. Write a **reference document** about DFT
    1. Example of prior work DCMI Metadata Terms
2. create an accompanying **abstract data organization model** that may be also expressed graphically,
3. **register the defined terms** in an ISO-like concept registry so that everyone can easily refer to them
    1. Proper data organization will be enabled by agreeing upon a number of basic concepts and their relationships as well by explicitly defining and registering appropriate terms along with alternative views of them
4. **engage many communities and stakeholders in the document, terminology and model creation**
5. establish **contact** with established, **relevant communities** such as W3C, librarians, etc.

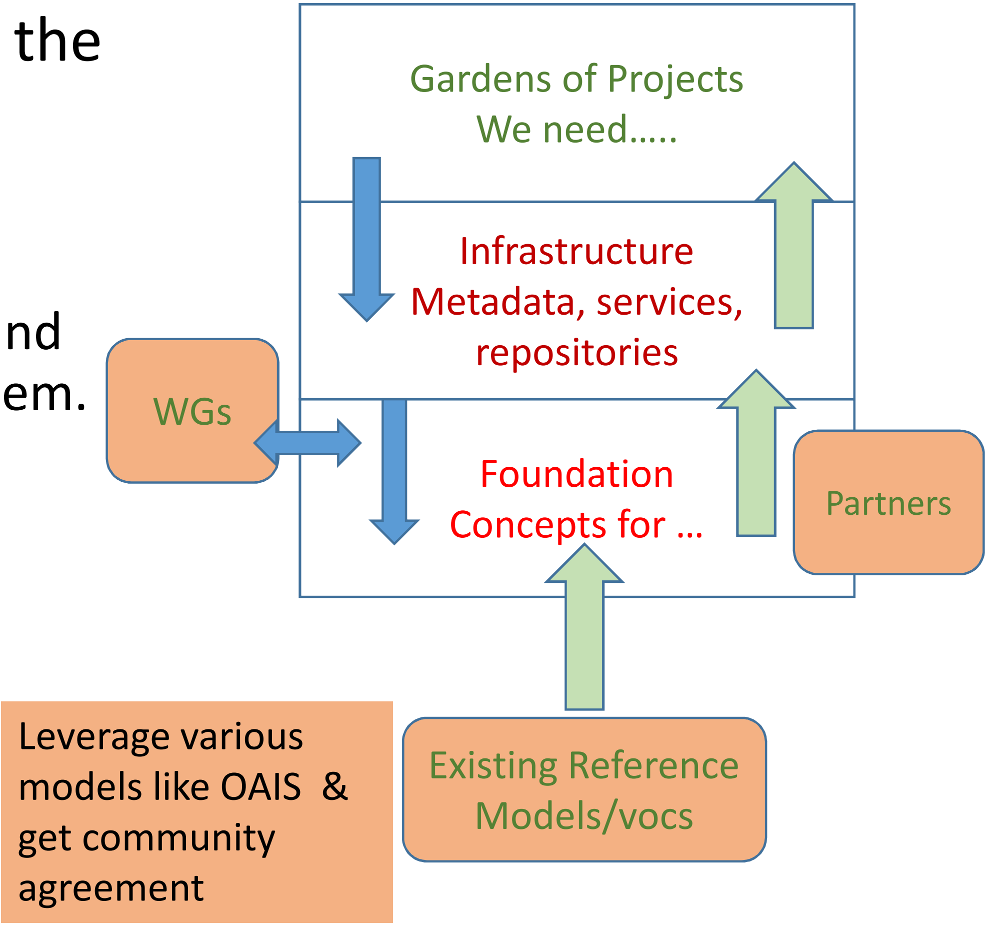
Dublin Core Example



Value Proposition of Shared Vocabularies

Research data communities helped by standardized data vocabularies reflecting the same definition for the same terms (data asset, data object, metadata….)

- help the RDA community to find common building blocks, describing their properties and defining data process protocols related to them.
  - Conversations/ interactions can be more meaningful
    - people aren't talking past each other.
  - Helps adoption of common data sharing practices and interoperation
- Help avoid duplication of effort.



Leverage various models like OAIS & get community agreement

Start Up and the following months

Startup Phase & Organizational activities	January 2013 - March 2013	Write a basic conceptualization note that describes the scope and intentions of the WG and
		Identify groups, initiatives, experts that worked on the issues and that are interested to participate in the work and motivate them to contribute to the discussion (in particular a broad involvement of data infrastructure projects is intended). This is of high priority for the group.
Phase 2	March 2013 - July 2013	Broadly discuss and extend the basic conceptualization note via the OIF and by having Video-Conference interactions. Instead of using the email-list more interaction will be done via the forum to indicate to others what is happening in the group. However there is a possibility for people to access the email-list exchange to inform themselves.
		Discuss this note in a f2f meeting in Gothenburg (March 2013)

Phases 3 and 4

Phase 3	August 2013 - December 2013	Consolidate the discussions with broader community involvement and converge on conceptualization and term definitions (Supported by face-to-face meeting in the summer)
		Start writing a reference document which is one of the outcome of the WG
		Define and register concepts in an ISO -like registry
		Interact with other RDA WG about terms used across working groups
Phase 4	January 2014 - April 2014	Write a final reference document
		Optimize the registered definitions
		Reach out and do dissemination and model roll out.

RDA 2nd Plenary: Discussion of Topics and Issues

- Simple vs. Complex Approach
  - Infrastructure design needs reduction, Abstractions… We need to define things as a way that we can work with them.
  - vs. Need to Understand what has been done
  - What is infrastructure?
  - For most concepts, there are no set definitions.
- Only part of life cycle covered in some models
- Document Decision Process
- How much existing vocabulary work are we leveraging and how?
- Concept-based terms or documenting existing terms?
- Are we being Foundational?
- Is it useful to define concepts like Metadata or just leave it?
- Some information is not included in the metadata, but in searchable data.
- Normative vs. non-normative parts of vocabulary (definition vs. examples)
- Proposal to use more elaborate frameworks used in vocabulary field
- What to communicate to other WGs?

Relevant, Engaged, Referenced work includes:

- Cross-disciplinary data infrastructure project EUDAT
    - iteratively using a proper conceptualization for a joint reference terminology helped boost subsequent discussions which lead towards solutions.
- ENVRI project (www.envri.eu), ongoing analysis and modeling
    - generic data organization and sharing model (in UML) covering: data acquisition, data curation, data access and data processing.
- analysis of 6 ESFRI infrastructures (acquisition, curation, access and processing)
- Kahn & Wilensky created a paper describing a framework for distributed digital object services.
- Wittenburg, Lautenschlager and Broeder analysed the data organizations of about 15 communities and came to some common abstractions.
- Concept models being developed as part of the Data Conservancy - http://dataconservancy.org/) Allen H. Renear David Dubin & others at Center for Informatics Research in Science and Scholarship (CIRSS) at U of Illinois

Adoption Plan

- Propose testing adoption on relevant projects:
  - Useful for other RDA WGs
  - Disciplinary EUDAT initiative and the DASISH cross-disciplinary initiative in the area of social sciences and humanities (Europe)
  - iCORDI *International Collaboration on Research Data Infrastructure* (US & Europe)
  - RENCI
  - Deep Carbon ….Others???
- Assumption is that many will get on board if reference model is inclusive and done well in response to input and involvement.
- Adoption is aided if work can support diversity through tools that employs a good representation usable by many if not everyone.
- Risc: being ignored (we've seen examples of that happening)

WG Membership

- About 40 members including people from other WGs
  - Some new people here???
  - More defined roles???
  - Liaison with EUDAT. What about others???
  - “It is about the alliance”
- Wide range of disciplines
- More from Europe than US or Australia
  - Still recruiting…..