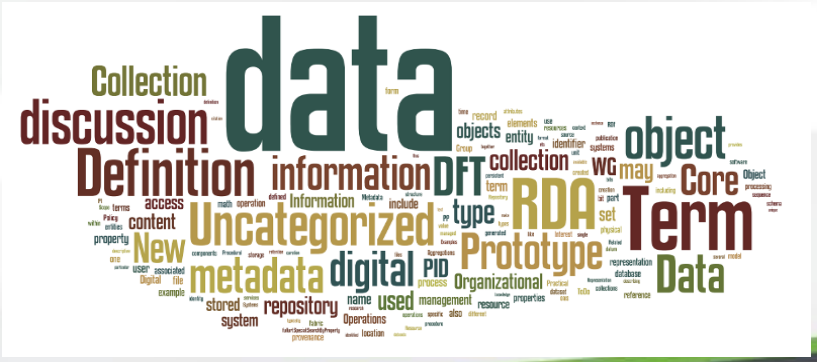
RESEARCH DATA ALLIANCE



Update on Data Foundations & Terminology (DFT) IG

https://rd-alliance.org/groups/data-foundations-and-terminology-ig.html

Gary Berg-Cross (SOCoP)

3rd Working Group Collaboration Meeting, June 11-12, 2015

Karlsruhe Institute of Technology

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.

Challenges: heterogeneous categorizations and varied methods for organizing information.

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

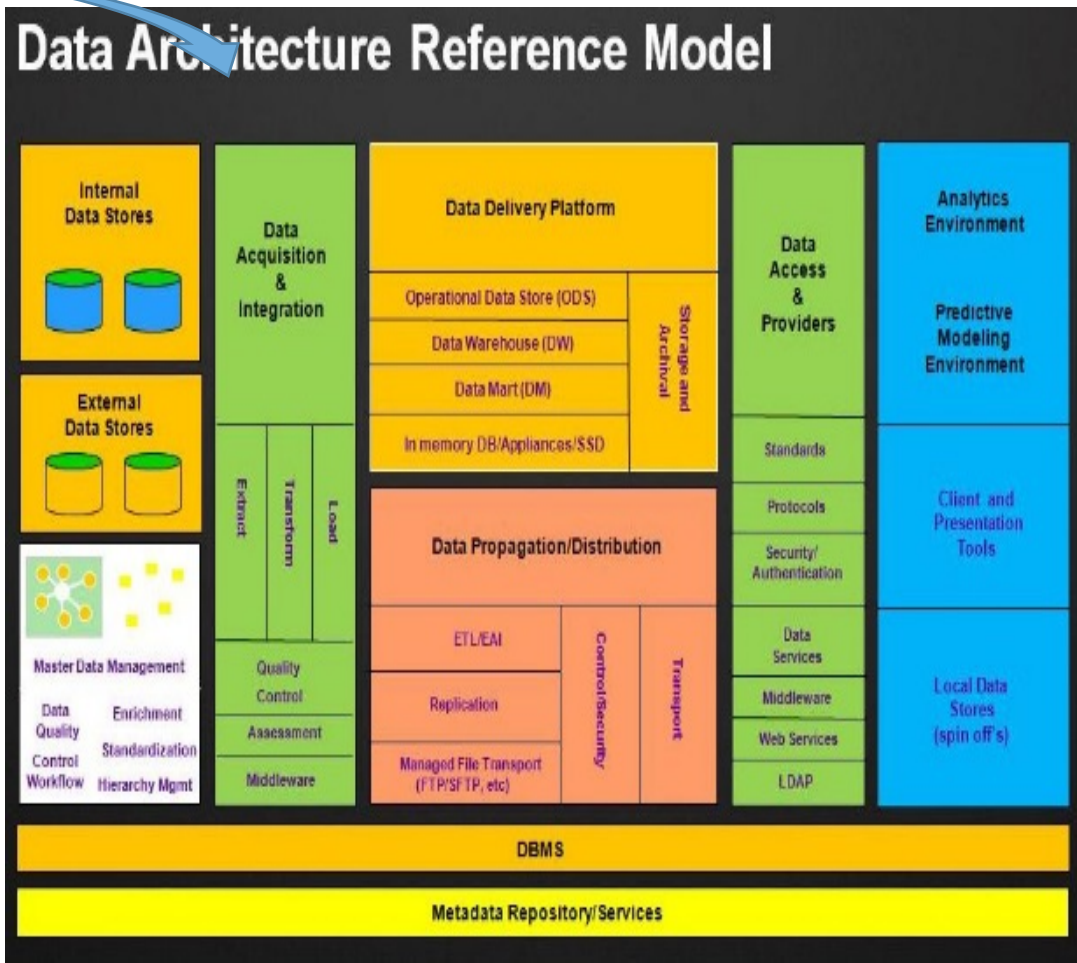
Infrastructure, Facilities, Operations

Acquisition
Storage
Preservation
Interoperability
Access
Services

Metadata
Identifiers
Curation
Policy
Discovery
Analysis
Tools

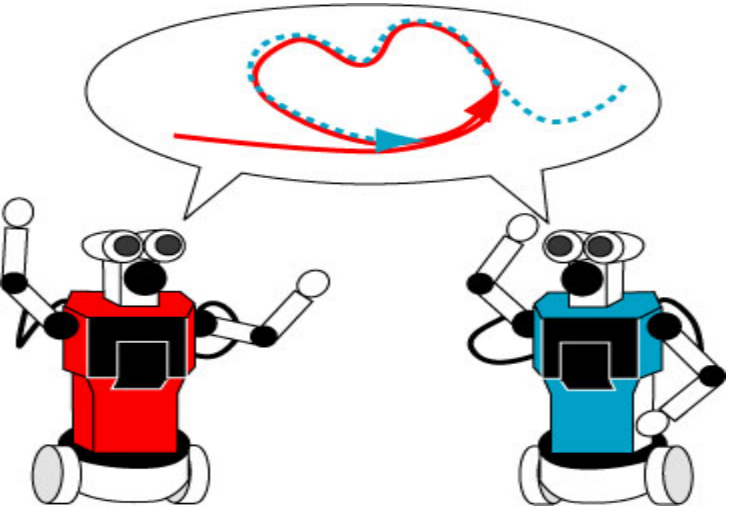
Software, Policies, Education

Discovery, Insight, Re-use



Shared Vocabularies are Needed for Understanding

Knowledge is a socially distributed phenomena.

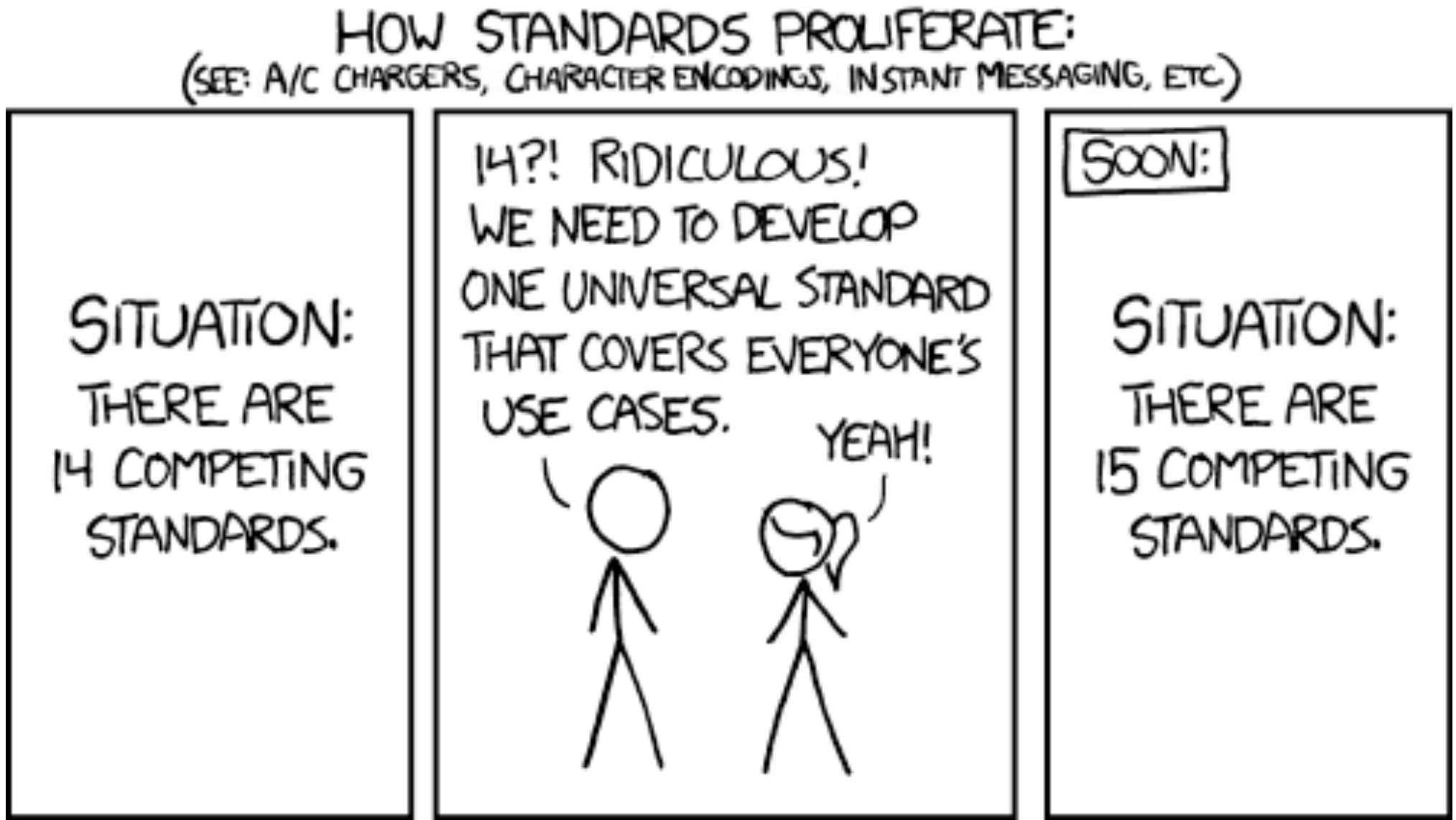


We build vocabularies to capture an understanding of knowledge in order to:

- facilitate **communication** across disciplines,
- **share/exchange data** and reuse it or
- enable **collaboration**

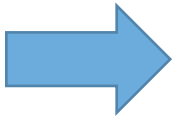
Trend is to leverage agreed upon conceptualizations & augment it with a formalized, digital representational of this knowledge allowing some degree of automated processing.

Something to Keep in Mind

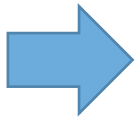


IG is Post DFT WG Activities

- Some progress describing some parts of an abstract 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.



Update as appropriate



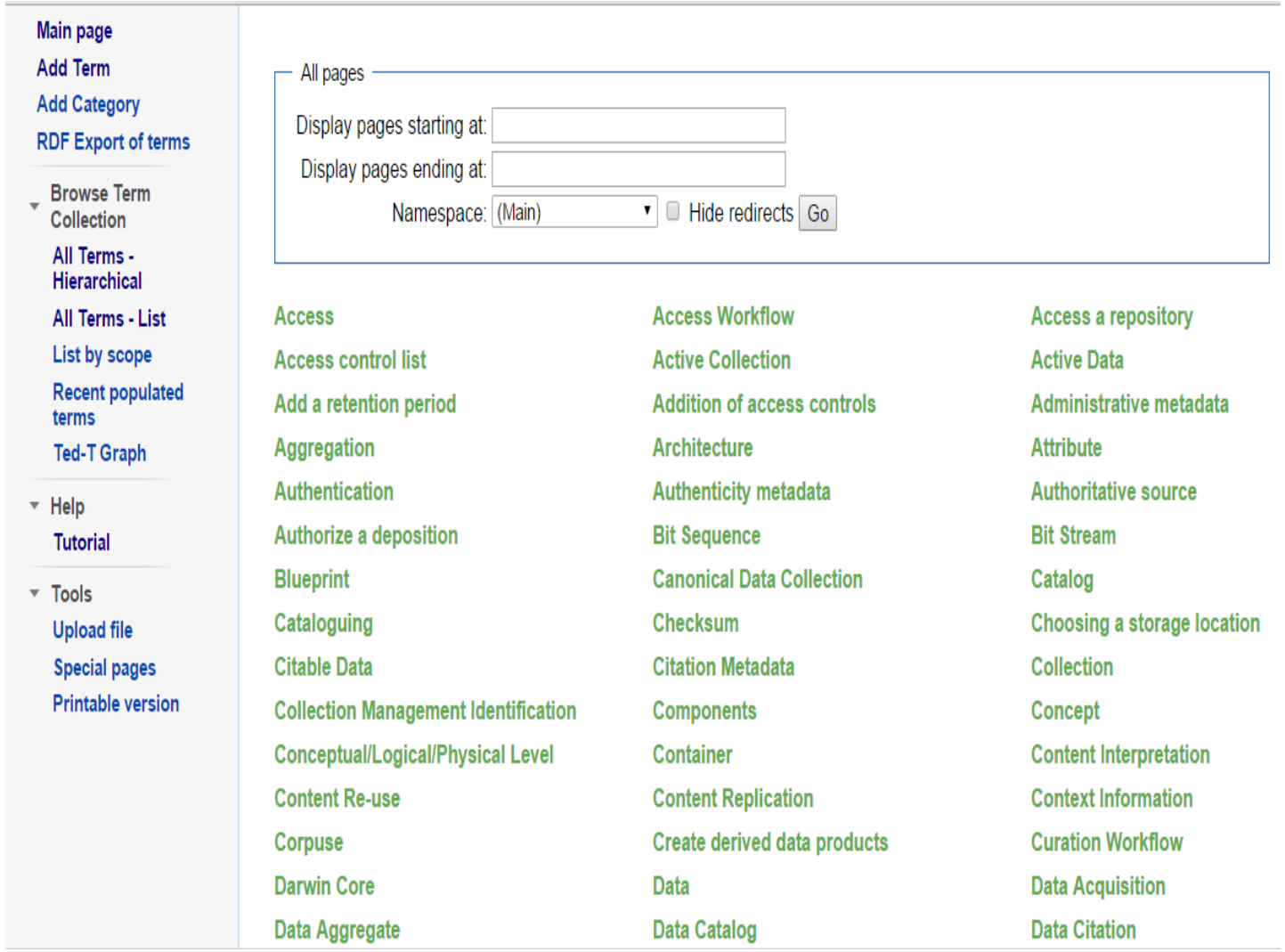
Populating 7 Updating Tool

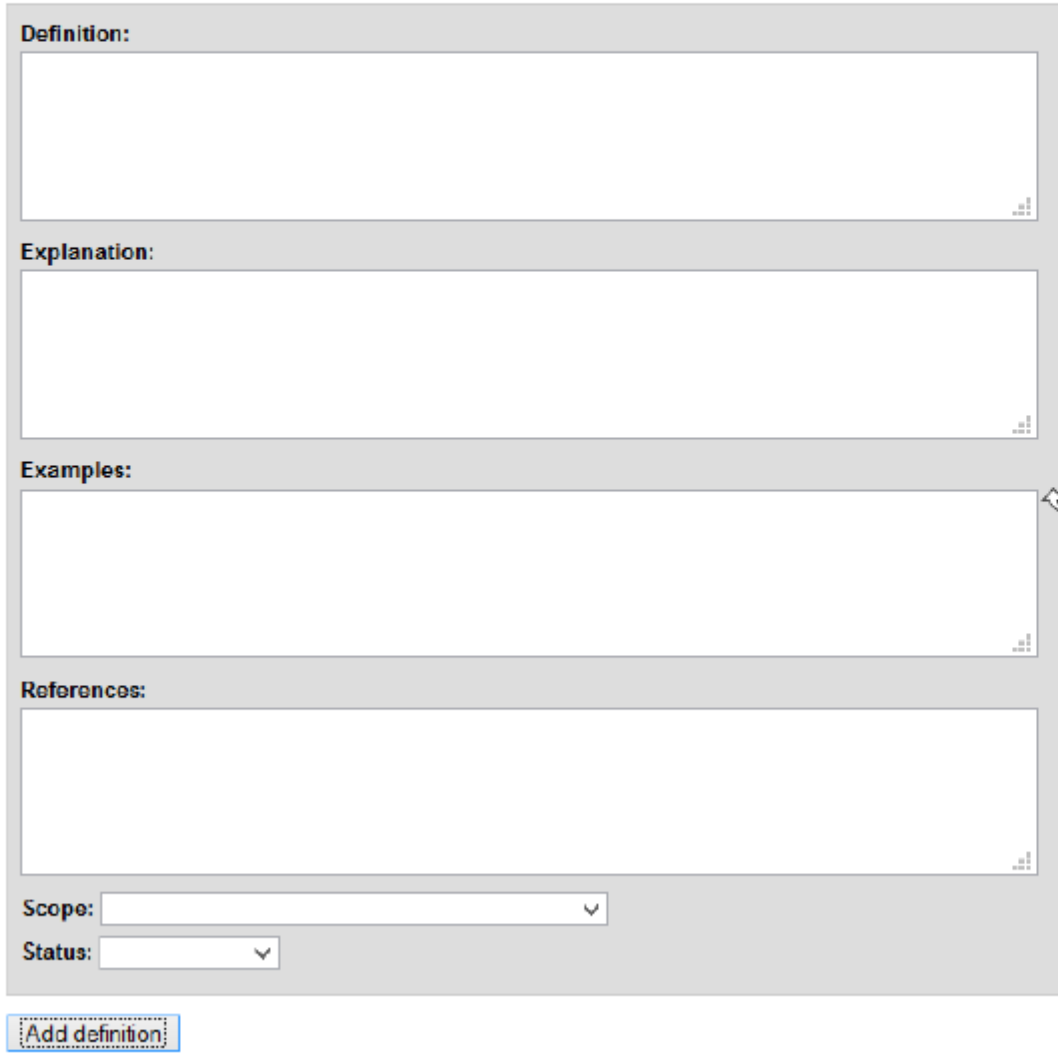
Future Work As IG 2015+

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 he 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.

Increasing # of Terms & Initial definitions are in TeD-T

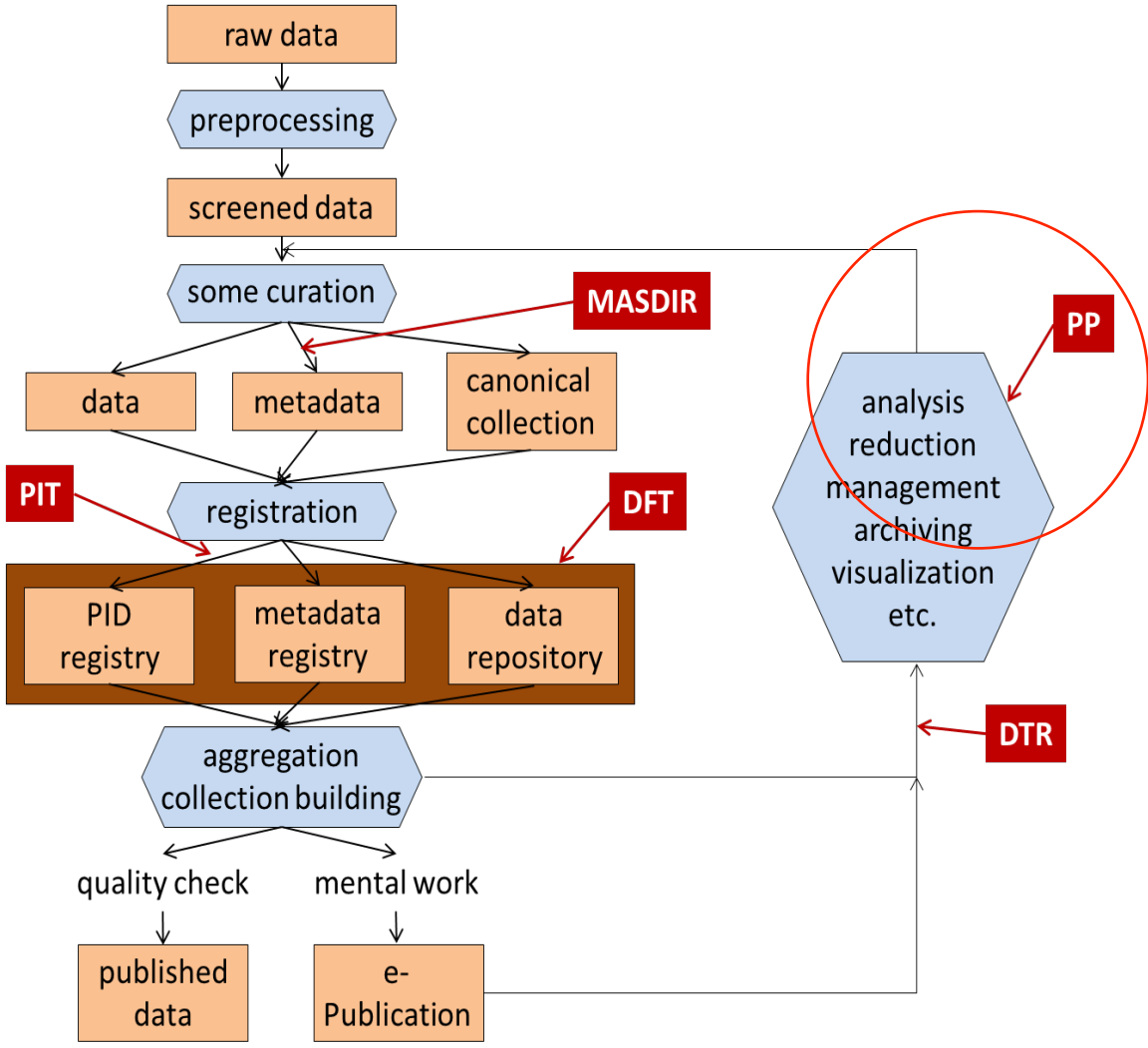




smw-rda.esc.rzg.mpg.de/index.php/Special:AllPages

Coordinating 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.**



Illustrating PP, MD View & Broader Scope

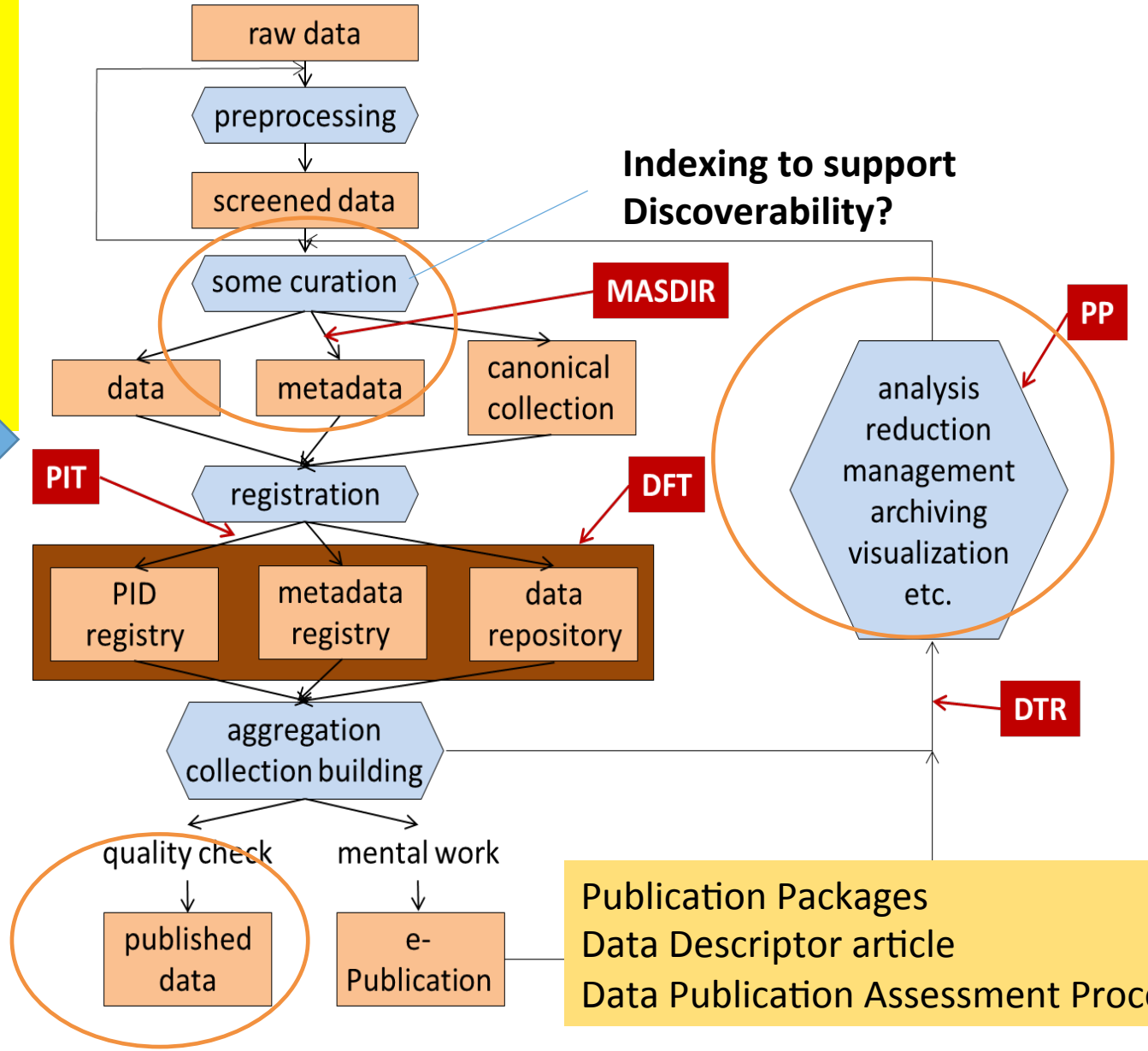
Policy defines this?

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?
E.g. Are registration & ingest the same thing?

Now Other WGs…..Linked Data?



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.

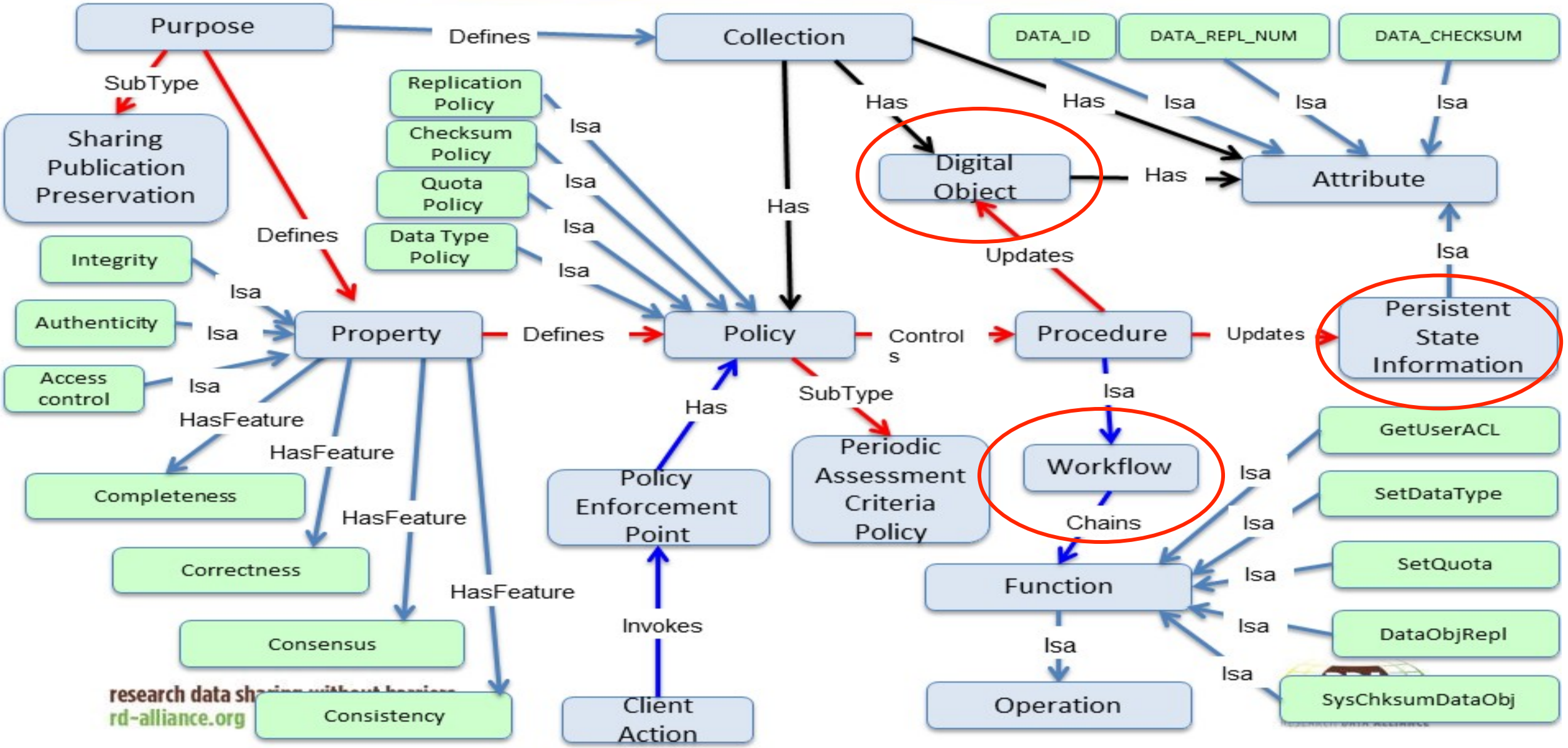
The creation of **provenance** and **descriptive** metadata defines a **context** for interpreting the relevance of files in a collection.
Depending upon the data source, there are multiple ways to provide metadata –**some automatable**:

- Extract metadata from an associated document. An example is the medical imaging format DICOM.

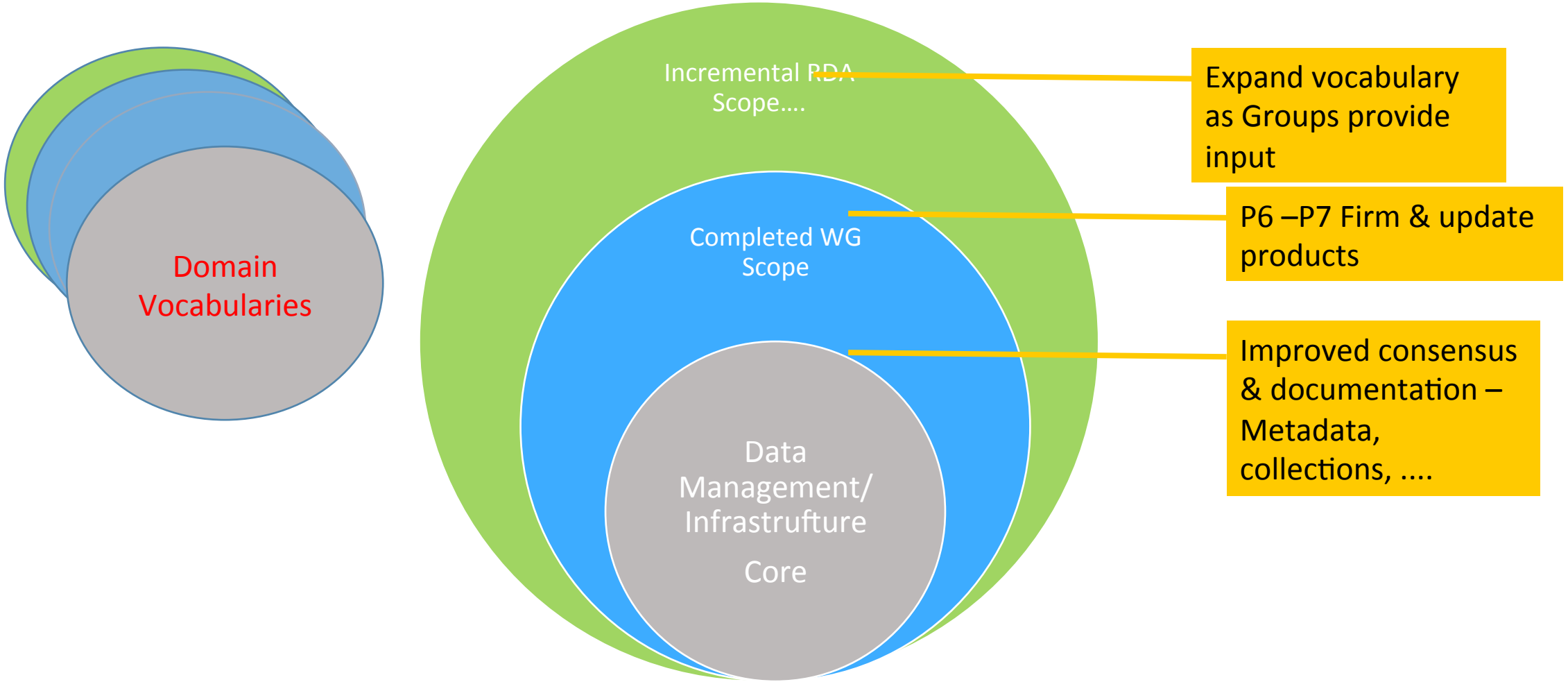
- Extract metadata from a structured document which includes **internal metadata**.
  - Examples are FITS for astronomy, netCDF, and HDF.
- Extract metadata by **parsing** patterns within the text within a document.
- **Identify a feature** present within a file and **label** the file with the location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

Policy Components - Conceptual Fundamentals

Policy-based Data Management Concept Graph



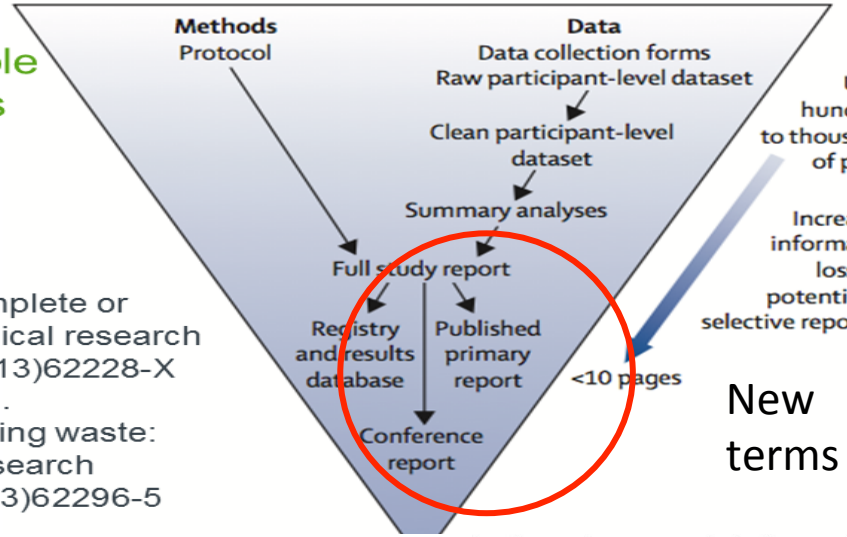
We can continue to Discuss & Built out Vocabulary in Stages –



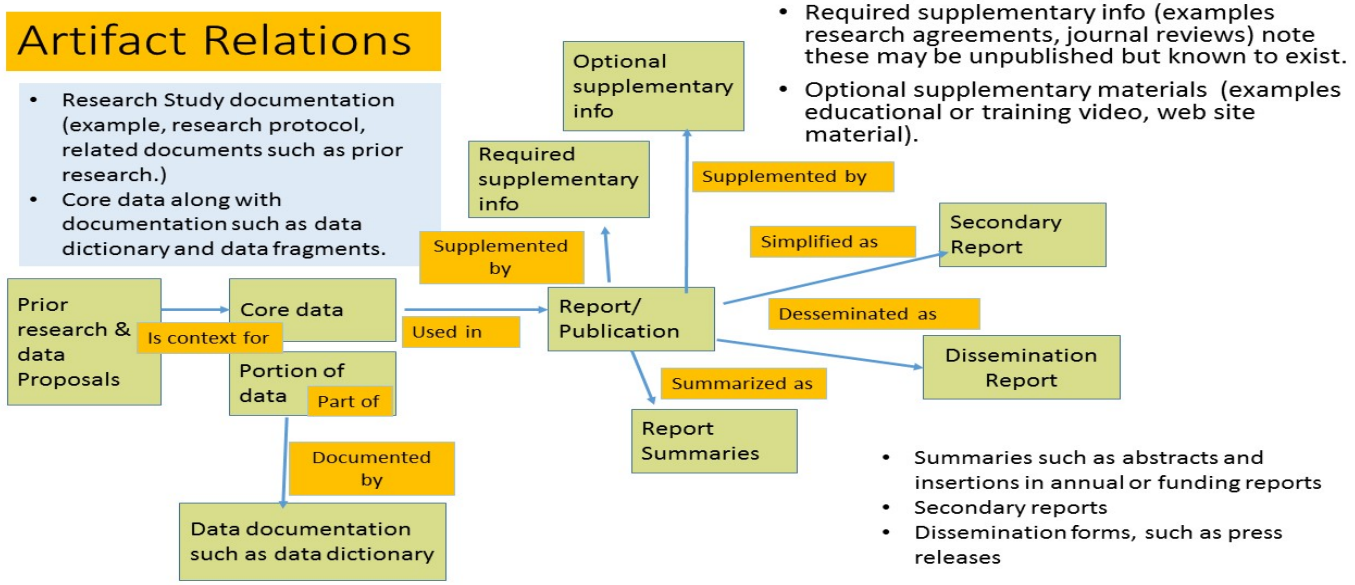
Use Cases Complex/Networked Research Objects

- Not all DOs are simple, solitary digital objects.
- "Earth System Science Data" (ESSD) Journal experience features, research data objects that are complex, compound and/or networked objects with many relations.
- Such compound objects are featured in the OAI-ORE model which is included in the DFT Model Overview paper.
- Publishing has several types and data publishing can occur at different stages of the research process.
- Adds concepts like „citation metrics“

- The Lancet, Jan 2014
- „9. Reliable and stable bidirectional linkages between all these elements“
- „9. …“: Paul Glasziou et al. Reducing waste from incomplete or unusable reports of biomedical research DOI:10.1016/ S0140-6736(13)62228-X
- Picture: An-Wen Chan et al. Increasing value and reducing waste: addressing inaccessible research DOI:10.1016/S0140-6736(13)62296-5

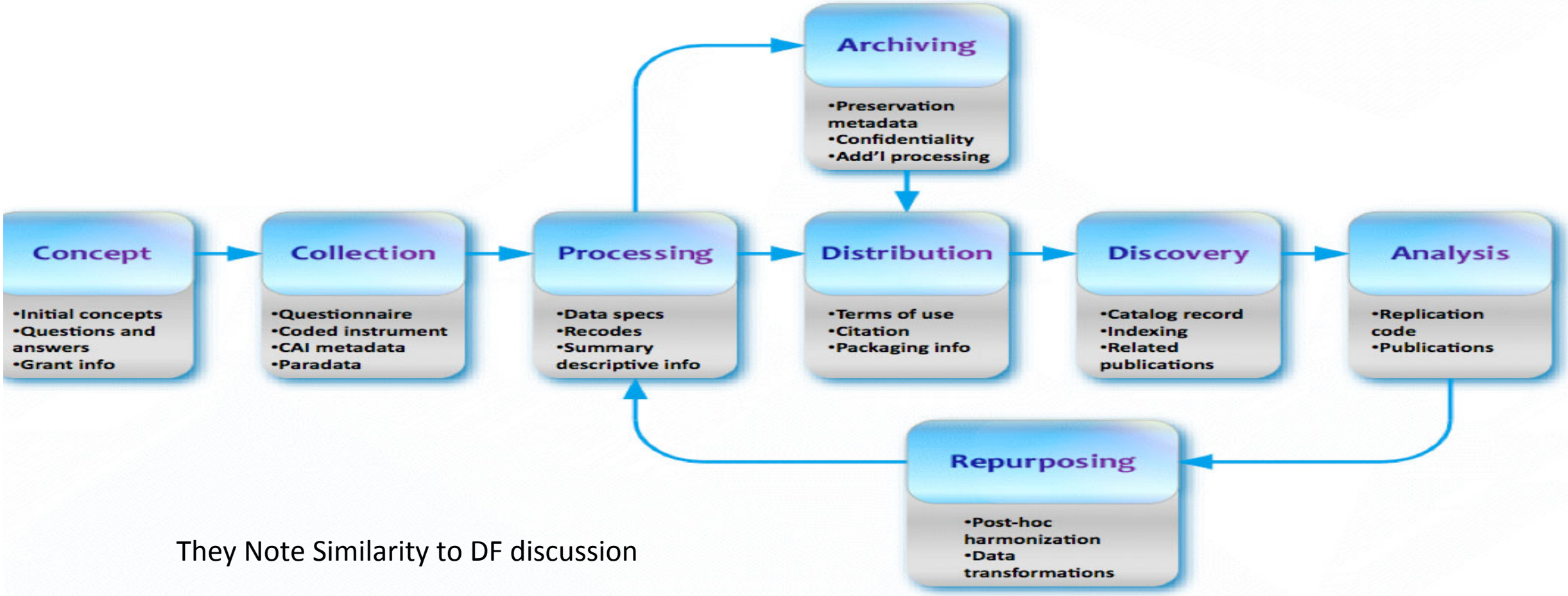


New terms



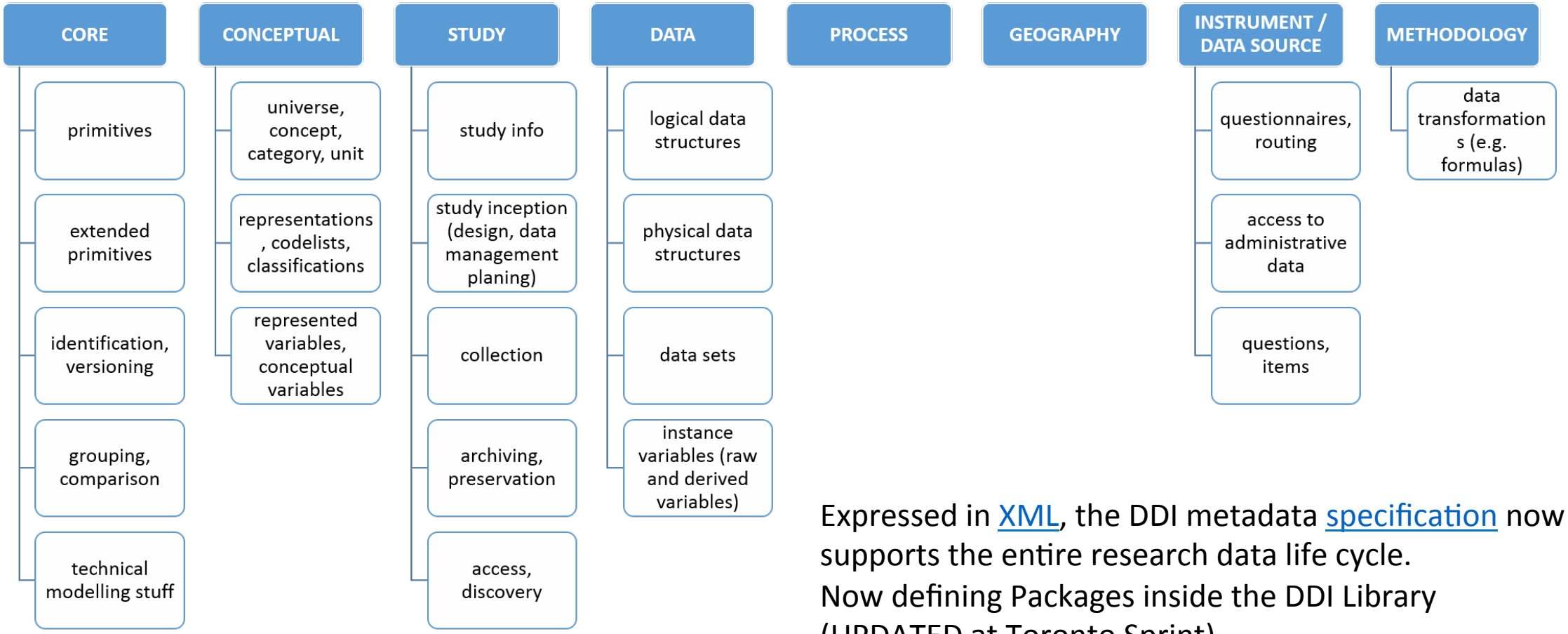
Outside Groups Are Interested in Collaboration:

Data Documentation Initiative (DDI) is an effort to create an international standard for describing data from the social, behavioral, and economic sciences.



They Note Similarity to DF discussion

DDI metadata accompanies and enables data conceptualization, collection, processing, distribution, discovery, analysis, repurposing, and archiving.



Expressed in XML, the DDI metadata specification now supports the entire research data life cycle.
Now defining Packages inside the DDI Library (UPDATED at Toronto Sprint)

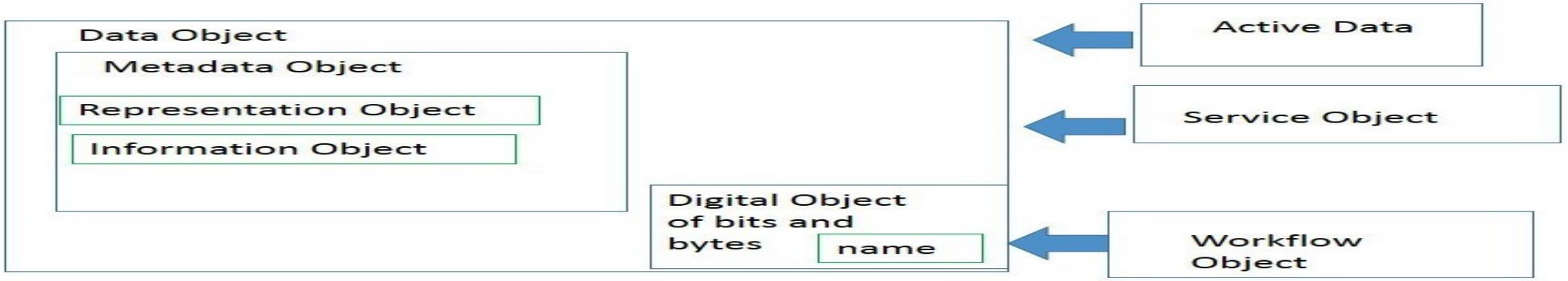
Lessons Learned and Follow Up in IG

- 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
- As part of an IG a broader plan for long-term maintenance & definitions upgrades
  - E.g. as metadata groups reach consensus
- A plan for term tool (TED-T) maintenance
  - updates for DFT terms and other WGs.

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

Simple Vocab Entry Example from P2 illustrates taxonomy & other relations, attributes etc.

- Data Object
  - Type of: Abstract Object (Taxonomy)
  - Sub-types: digital object,……
- Definition: n 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)
  - http://en.wikipedia.org/wiki/Data_object
  - 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)

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)

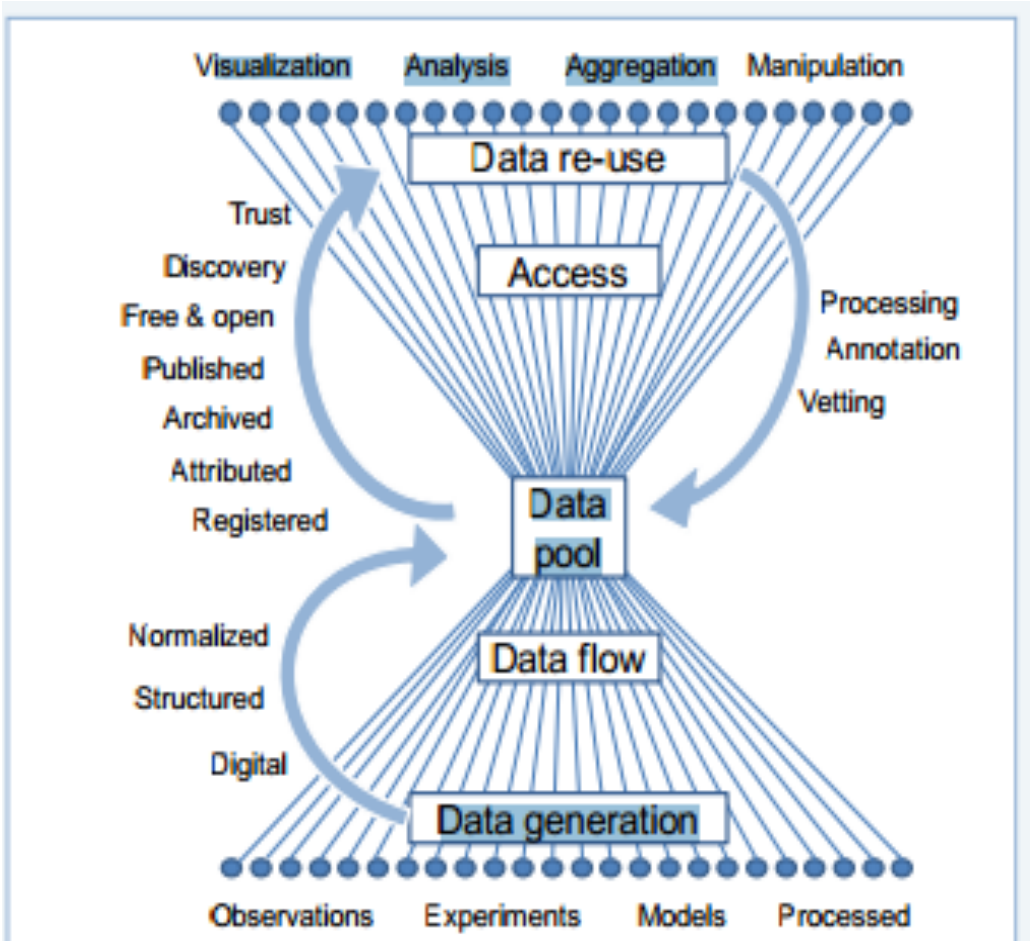


Figure 8: A Big New Biology can only emerge with a framework that optimizes reuse. Ideally, data should be in forms that can flow from source into a common pool and can flow back out to consumers, be subject to quality control, or be enhanced through analysis to rejoin the pool as processed data.

Notional Core Diagram Reflecting Data Lifecycle and RDA WGs