

Thematic Clustering of RDA Working and Interest Groups

Proposal to RDA Technical Advisory Board

V2. 18 Jan 2015

Beth Plale, Peter Fox, Francoise Genova, Inna Kouper, Rainer Stotzka, Peter Wittenburg

I. Motivation

RDA Plenary 4 was a tremendous success. Its 500 participants and considerable activity spoke to the timeliness and relevance of RDA and its efforts. At the same time, RDA leadership: TAB, Council, OAB, and Secretariat, heard repeatedly that RDA is difficult to comprehend. Plenary attendees had difficulties recognizing focus or path in the activities of the more than fifty Working and Interest Groups (WGs / IGs).

TAB began discussion of area clustering its post P4 TAB meeting, and there was clear consensus existed that something needed to be done and soon. The early ideas behind the proposed clustering emerged from a back-of-napkin discussion at the WG/IG meeting in Washington DC Nov 2014, with Beth Plale, Kathy Fontaine, Jay Pearlman, and Francoise Pearlman. Mark Parsons then developed the notions further in front of the WG/IG group. A small sub-group of TAB members plus Engagement IG co-chair, Inna Kouper formed and met winter of 2014-15 to formulate this proposal document.

This document proposes a thematic clustering of activity that maximizes the commonality among WG/IG activities. It further identifies and limits the purposes of the thematic clustering.

The purpose of WG/IG clustering is several-fold:

1. Guide **newcomers** in finding knowledge, expertise, and solutions and in joining appropriate groups.
2. Help **externals** to find focus and coherence of RDA's approach and solutions.
3. Guide **RDA members** who want to start a new activity in what is already being done and how to avoid overlaps.
4. Inform **WG/IG members** about other groups' activities.
5. Help **TAB** in guidance and evaluation of existing and new groups.
6. Help **TAB and other coordinating bodies** to identify gaps and overlaps in describing RDA and determining a "roadmap".

It should be specially noted that clustering does not obligate WGs/IGs and their chairs to meet or work together unless they voluntarily decide to do so.

Next Steps:

- TAB agreement – end of January
- Council acknowledgement – end of February
- Roll-out at P5

II. Methodology

The following principles guided the development of the proposed clustering:

1. Intuitiveness. The criteria for classification must be intuitive and follow the practices of the communities.
2. Flexibility. The criteria and WGs/IGs assignments can be revised at all moments dependent on the experience, no one system is fully satisfactory.
3. Sensitivity. The choice of areas must be sensitive to the full suite of policy, legal, and technological breadth across RDA.
4. Manageable size. Each cluster should have no more than 10-15 members.
5. Ease of navigation. One has to be able to browse, search and filter on multiple dimensions.

Several approaches to clustering have been considered, including the data lifecycle stages approach, functions in phases approach, WG/IG collaboration workshop taxonomy, and word frequency approach (see Appendix A for details). While those approaches offer several useful dimensions for clustering, none of them could be used alone and satisfy the principles outlined above. Therefore, we draw on elements from all four approaches and propose a hybrid thematic approach below.

III. Thematic Clustering

The clustering is done along the two dimensions: a *solution dimension* (Y-axis) and the *beneficiary dimension* (X-axis) where the solution dimension is a spectrum from technical to social, and the beneficiary dimension is a spectrum from data providers to users. Each Working Group and Interest group occupies a single point in this “Cartesian space”. It was pointed out at the Nov 2014 WG/IG meeting that groups may actually have non-point, or polygon representations in this space, but that is left for a refinement.

The Cartesian space when divided into four quadrants yields the following quadrant descriptions. In categorizing interest groups (IGs) into the four quadrants, common terms emerge from the IGs that are collocated; these are given as well.

Q1: Social/educationally oriented activity that benefits users

Common terms: education, engagement, bridging, community

Q2: Technically-oriented solutions that benefit users

Common terms: interoperability, harmonization, integration, metadata

Q3: Technical solutions that aid in data provisioning

Common terms: repository, fabric, analytics, identity, management

Q4: Policy oriented solutions that aid in data provisioning

Common terms: governance, certification, cost recovery, legal

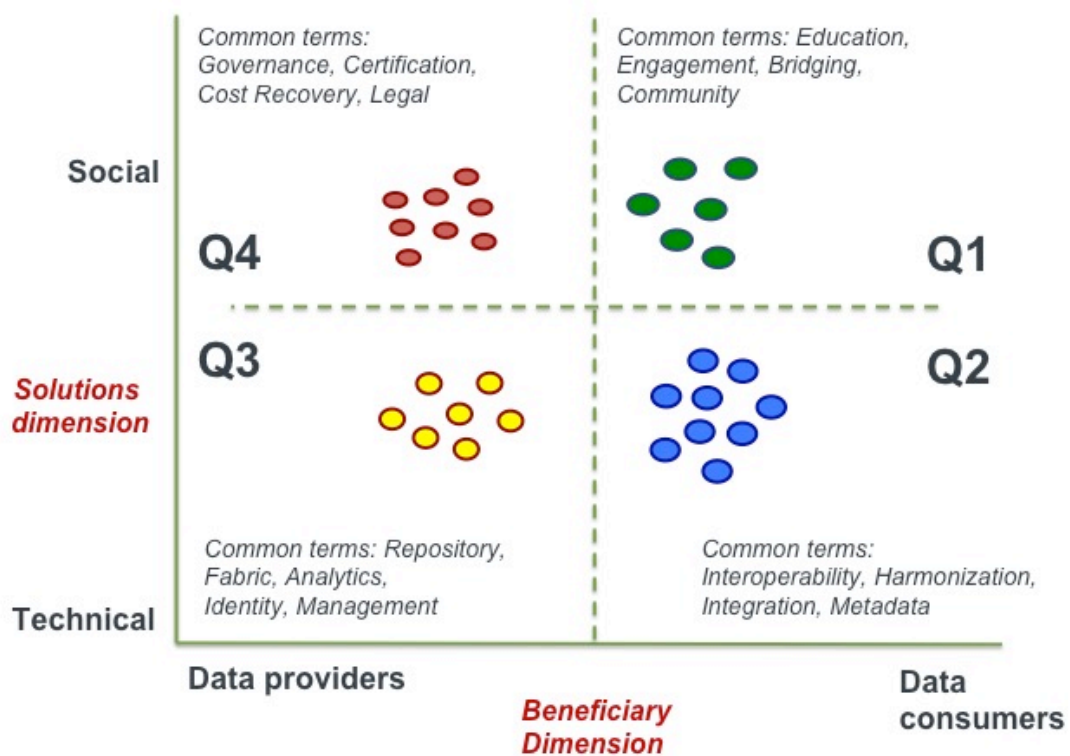


Figure 1. Solution-beneficiary thematic clustering approach. The common terms are taken from group names in the cluster, and used here to give meaning to the quadrant.

a. Thematic Clustering – Working Groups

Working Groups are arranged in the Cartesian space as shown in Figure 2. Interest Group and Working Group breakout is shown in table form in Table 1. The commonly occurring terms used in Figure 1 are underlined there.

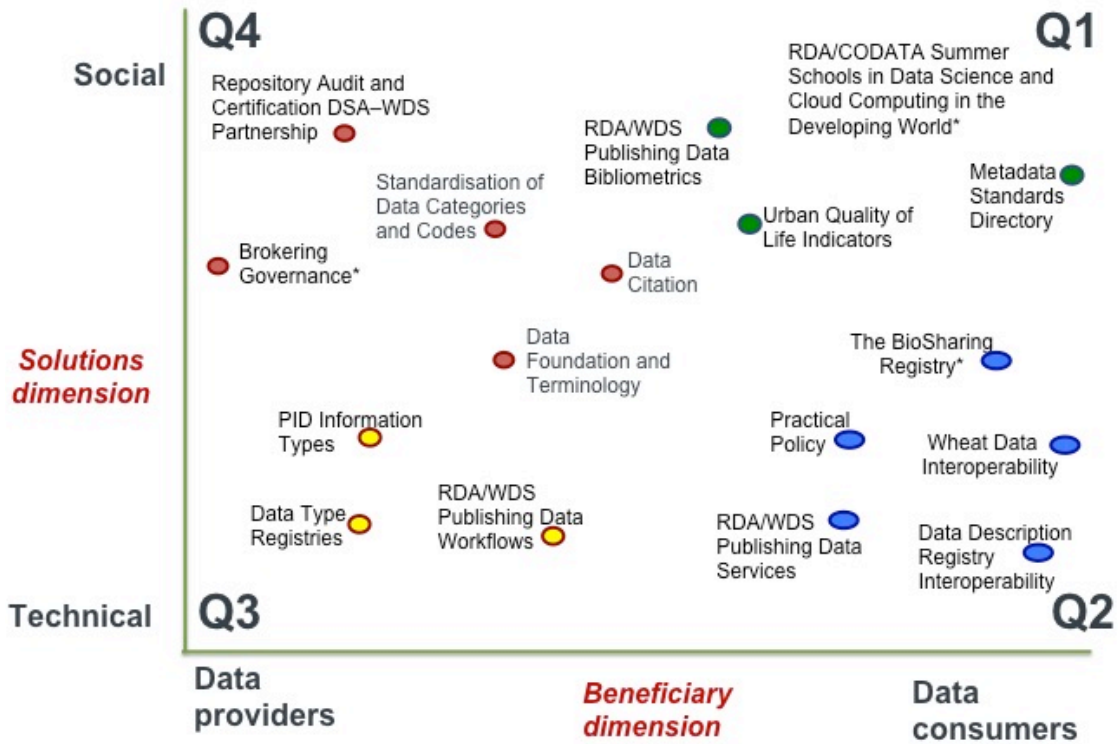






Figure 2. Working Groups Clusters

IV. Tagging – As Way to Further Describe

In addition to clusters, we propose the use of terms by which a WG/IG self-identifies. These terms can further categorize groups and aid navigation. A preliminary list of tags is; tags are added by WG/IG groups as needed:

| | | | |
|------------------|---------------|-----------------|--------------|
| Education | Libraries | Data Discovery | Preservation |
| Governance | Data Modeling | Data Fabric | Protocols |
| Interoperability | Networks | Data Publishing | Big Data |

Table 1. Interest Groups (blue) and Working Groups (brown) Cluster by Group

| Q1 Social/educational activity in aid of data consumers  | Q2 Technical solutions in aid of data consumers  | Q3 Technical solutions in aid of data provisioning  | Q4 Policy solutions in aid of data provisioning  |
|--|--|---|--|
| Community Capability Model (CCM) | Agricultural Data Interoperability | Big Data Analytics | Brokering Governance |
| Development of Cloud Computing Capacity and Education in Developing World Research | Biodiversity Data Integration | Data Fabric | Digital Practices in History and Ethnography |
| Education and Training on Handling of Research Data | Geospatial | Data in Context | RDA/CODATA Legal Interoperability |
| ELIXIR Bridging Force | Marine Data Harmonization | Domain Repositories | RDA/WDS Certification of Digital Repositories |
| Engagement | Metabolomics | Federated Identity Management | RDA/WDS Publishing Data Cost Recovery for Data Centres |
| Libraries for Research Data | Metadata | Persistent Identifiers | RDA/WDS Publishing Data |
| Long Tail of Research Data | RDA/CODATA Materials Data, Infrastructure & Interoperability | Preservation e-Infrastructure | Research Data Provenance |
| Research Data Needs of Photon and Neutron Science community | Structural Biology ?? | PID Information Types | Service Management ?? |
| Development of Cloud Computing Capacity and Education in Developing World Research | Toxicogenomics Interoperability | Data Type Registries | Brokering Governance |
| RDA/CODATA Summer Schools in Data Science and Cloud Computing in Developing World | Agricultural Data Interoperability | RDA/WDS Publishing Data Workflows | Digital Practices in History and Ethnography |
| Urban Quality of Life Indicators | RDA/WDS Publishing Data Services | | Repository Audit and Certification DSA-WDS Partnership |
| Metadata Standards Directory | Practical Policy | | Brokering Governance |
| | Wheat Data Interoperability | | Standardization of Data Categories and Codes |
| | Data Description Registry Interoperability | | Data Foundation and Terminology |
| | BioSharing Registry | | Data Citation |
| | | | |

Appendix A.

Data Lifecycle Stages Approach

The Data Lifecycle stages approach can be used to cluster groups based on their focus relative to the stages that data go through, e.g., the stages of collection, analysis, and preservation. Figure 4 below adapted from the DataOne project¹ and extended by adding the stage “Publish” illustrates all the stages.

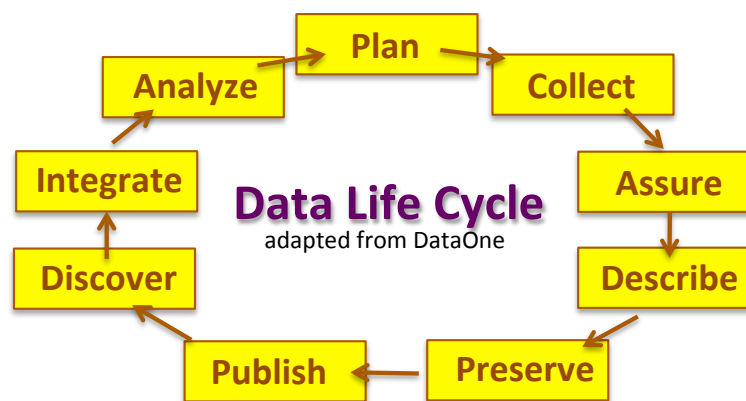


Figure 3. Data Lifecycle.

The table below provides an example of how the existing WGs can be mapped into the data lifecycle stages.

| Working Group | Plan | Collect | Assure | Describe | Preserve | Publish | Discover | Integrate | Analyze |
|---|------|---------|--------|----------|----------|---------|----------|-----------|---------|
| Brokering Governance | | | | | | | x | x | x |
| Data Description Registry Interoperability | | | | xx | | x | x | | |
| Data Foundation and Terminology WG | x | x | x | x | x | x | x | | |
| Data Type Registries WG | | | x | x | | | xx | x | x |
| Metadata Standards Directory WG | x | x | | xx | x | x | x | | |
| PID Information Types WG | | | | xx | x | x | x | x | x |
| Practical Policy WG | | x | x | x | x | x | x | x | x |
| RDA/CODATA Summer Schools in Data Science and Cloud Computing in the Developing World | | | | | | | | | x? |

¹ See <https://www.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf>

| | | | | | | | | | |
|---|--|--|----|----|---|----|---|--|--|
| RDA/WDS Publishing Data Bibliometrics WG | | | | | | XX | | | |
| RDA/WDS Publishing Data Services WG | | | | | | XX | | | |
| RDA/WDS Publishing Data Workflows WG | | | | | | XX | | | |
| Repository Audit and Certification DSA–WDS Partnership WG | | | XX | X | X | | | | |
| Repository Platforms for Research Data | | | X | X | X | X | | | |
| Standardisation of Data Categories and Codes WG | | | | X? | | | | | |
| The BioSharing Registry: connecting data policies, standards & databases in life sciences | | | | | | XX | | | |
| Urban Quality of Life Indicators | | | | X | X | | X | | |
| Wheat Data Interoperability WG | | | | | | | | | |

Functions in Phases Approach

The diagram below depicts functional phases of activities associated with data, such as data collection, registration, processing, storage and publication. For several groups it is easy to assign them to phases, some are relevant for a number of phases and some are relevant across almost all phases.

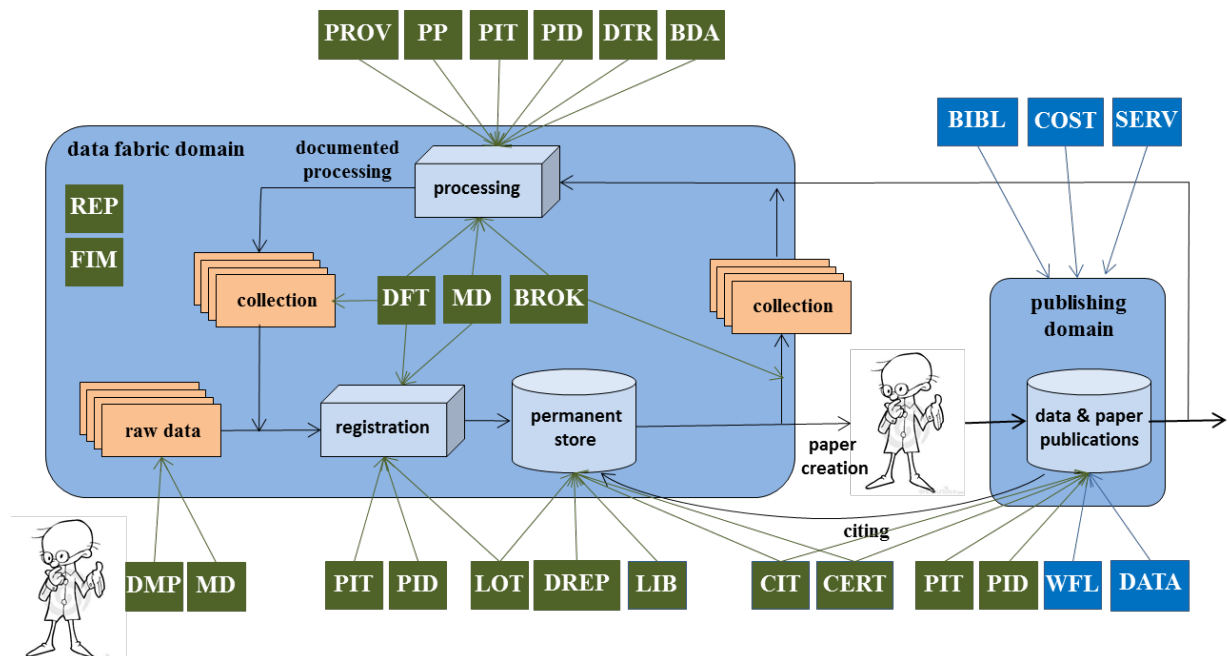


Figure 4. Functional phases of data; the following abbreviations are used in the diagram: **Brok**: Brokering WG & IG, **CIT**: Data Citation, **DFT**: Data Foundation & terminology, **DTR**: Data Type Registries, **MD**: Metadata WG & IGs, **PIT**: PID Information Types, **PP**: Practical Policies, **CERT**: Repository Certification, **DMP**: Active Data Management Plans, **BDA**: Big Data Analytics, **PROV**: Research Data Provenance, **REP**: Reproducibility, **DREP**: Domain Repositories, **FIM**: Federated Identity Management, **LIB**: Libraries for Research Data, **LOT**: Long Tail Data

The WG/IGs that have a direct link to the Data Fabric are colored green. The WGs/IGs that focus on publication aspects are in blue: BIBL, COST, SERV, WFL, DATA. As this attempt shows, groups that are not focused on functional phases of data are more difficult to fit into this diagram.

WG/IG Collaboration Workshop Taxonomy

This grouping was discussed at an RDA WG chairs meeting in Munich in 2013 and was widely agreed upon. In the table below, groups are organized according to topics. The last column also assigns layers, which are described in another table below.

| cat1 | cat2 | cat3 | WG/IG | WG/IG Topic | Layer | |
|----------------------------|---------------------------|-------------------------------------|------------------------------------|--|--|-------|
| cross-disciplinary Groups | technical | Semantics | WG | Data Foundation and Terminology | D/E | |
| | | | | Standardisation of Data Categories and Codes | D/E | |
| | | identifiers / referring | IG | Semantic Interoperability | D/E | |
| | | | WG | PID Information Types | B | |
| | | metadata | WG | | PIDs | B |
| | | | | | Metadata Standards Directory | A/D/E |
| | | | IG | | Data Description Registry Interoperability | A |
| | | | | | Research Data Provenance | A/D/E |
| | | registry workflow/ processing | WG | | Data in Context | A/D/E |
| | | | | | Metadata | A/D/E |
| | | | WG | | Data Type Registries | D |
| | | | | | Practical Policy | E |
| | | | IG | | Big Data Analytics | E |
| | | | | | Long tail of research data | E |
| | Repository/ Federating | IG | | Brokering | I | |
| | | | | Federated Identity Management | I/C | |
| | | | | Preservation e-Infrastructure | G/H | |
| | non-technical | publishing/ citation | WG | Data Citation | A | |
| | | | IG | Publishing Data | A | |
| | | quality | IG | Certification of Digital Repositories | G/H | |
| legal | | IG | Legal Interoperability | C | | |
| community | | IG | | Community Capability Model | X | |
| | | | | Development of cloud computing capacity and education for developing world | X | |
| | | | | Engagement Group | X | |
| discipline-specific groups | agriculture | WG | Wheat Data Interoperability | X | | |
| | | IG | Agricultural Data Interoperability | X | | |
| | biology | IG | | Toxic genomics Interoperability | X | |
| | | | | Structural Biology | X | |
| | | | | Biodiversity Data Integration | X | |

| | | | | | |
|--|-------------------|--|----|--|---|
| | environment | | IG | Marine Data Harmonization | X |
| | Humanities/SocSci | | IG | Defining Urban Data Exchange for Science | X |
| | | | | Digital Practices in History and Ethnography | X |

Layers codes description:

| Functional Access and Management Layers | |
|--|---|
| Find/Reference | A |
| Ref-Resolution | B |
| Access | C |
| Interpret | D |
| Re-use/process | E |
| Manage | F |
| Curate | G |
| Archive | H |
| Federate | I |

Affinity by Word Frequency

An affinity approach was done in late 2014 based on word frequency analysis and qualitative coding of the wikis and web pages of each RDA group. It was performed by Candice Lanius. While this approach generates too many clusters to navigate through, some affinities can be used as additional categories that supplement the primary clustering.

1. Brokering Governance WG, Brokering IG, RDA/CODATA Legal Interoperability IG, and Service Management IG. Logic: Each of these groups is invested in bridging existing, large scale, international infrastructures. Brokering and federated services pose technical solutions and problems that intersect with discussions of the legal interoperability of research data.
2. Service Management IG, and Federated Identity Management IG. Logic: The Federated Identity Management (for authentication and authorization across platforms) is one component of the Service Management's interest in shared service delivery and data infrastructures.
3. Data Citation WG, Publishing Data Workflows IG, and Publishing Data IG. Logic: Publishing issues from the researcher's perspective.
4. Data Foundation and Terminology WG (and IG), and Community Capability Model IG. Logic: These groups look at data sharing issues at the organizational level. From an ideal abstract

description of use cases, services/ tools, and infrastructure to the capability models which look at the gaps in real world organizations and domains.

5. Data Type Registry WG, Standardization of Data Categories and Codes IG, (Big Data Analytics IG). Logic: These groups are invested in determining a set of core terms and common language for data use and management.
6. Metadata Standards Directory WG, PID Information Types WG, Metadata IG, PID IG. Logic: The creation of permanent ways to track the contextualizing information for data sets.
7. Summer Schools in Cloud Computing WG, Development of cloud computing capacity and education in developing world research IG, Education and Training on handling research data IG. Logic: Share information about developing curriculum and managing the logistics of courses.
8. Publishing Data Services WG, Publishing Data Bibliometrics WG, Repository Platforms for Research Data IG, Domain Repositories IG, (Publishing Data Cost Recovery for Data Centres IG). Logic: Publishing and data management from the perspective of service providers.
9. Repository Audit and Certification WG, Preservation e-Infrastructure IG, Certification of Digital Repositories IG. Logic: Preservation e-infrastructure is interested in expanding capabilities, which aligns with the knowledge and expertise of the repository certification groups.
10. The BioSharing Registry IG, Biodiversity Data Integration, Metabolomics IG, Structural Biology IG, and Toxicogenomics IG. Logic: Domain specific.
11. Digital Practices in History and Ethnography IG, Engagement IG. Logic: A unifying interest in ethnography of RDA practices and culture.
12. Urban Quality of Life Indicators IG, Geospatial IG, Data for Development IG, (Digital Practices in History and Ethnography IG). Logic: New ways to handle qualitative data across domains.
13. Wheat Data Interoperability WG, Agricultural Data IG. Logic: Domain specific.
14. Active Data Management Plans IG, Data in Context IG, Research Data Provenance IG. Logic: All of these groups are interested in establishing and maintaining data provenance/ context, with the management plan being a dynamic response to changing circumstances.
15. Libraries for Research Data IG, Long tail of research data IG, Logic: University specific data archiving and the interests of research libraries.

Groups without clear matches:

- Data Description Registry Interoperability WG
- Practical Policies WG
- Research Data Needs of Photon and Neutron Science Community IG
- Materials Data, Infrastructure & Interoperability IG
- Marine Data Harmonization IG

Umbrella Groups:

- Data Fabric IG
- Metadata IG
- Ethics and Social Aspects of Data IG
- Reproducibility IG

Common Topics

- Use-Cases
- Curriculum/ Education
- Qualitative Data
- Big Data
- Data Repositories
- Metadata
- Context/ Provenance
- Business/ Funding
- Publishing
- Service/ User Agreements/ Federated Management
- Data Management