

MOU to: RDA WG and IG Chairs

Subject: RDA Working Group / Interest Group Cluster Model for Thematic Areas

Action: TAB seeks feedback from RDA WG and IG Chairs on two pieces of the area cluster model described in this document.

Due by: 27 Feb 2015

After considerable effort, community input, and deliberation, TAB has endorsed an area cluster model that we think will bring clarity to the RDA activity particularly for the external community. TAB now seeks feedback from RDA WG and IG Chairs on two pieces of the area cluster model described in this document. The two pieces of feedback are:

1. After reading over the area cluster model below that TAB has endorsed. Look over where your WG or IG is located in the quadrant clustering of Figure 2 (for WGs) and Table 1 (for WGs and IGs). The categorization was done by your TAB liason(s) to the best of their knowledge, but you as chair are the ultimate expert. Send your feedback by email to contacts below.
2. Read Section IV on tagging. Select a set of tags 3-5 that best describe your WG or IG activity. Tags are drawn from the table, or can be of your suggestion.

Send your feedback to Beth Plale and Françoise Genova who are TAB's designated point people for feedback and who will aggregate results, analyze, and incorporate. Email addresses of both are provided in the email though which you received this MOU.

TAB is eager to hear from you. Please respond by 27 February 2015, so we can reflect your feedback before P5.

I. Introduction

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 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 50+ Working and Interest Groups (WGs / IGs).

History: Several approaches to clustering have been considered over the last 12 months, including the data lifecycle stages approach, functions in phases approach, WG/IG collaboration workshop taxonomy, and word frequency approach (these are in appendix).

TAB began discussion of area clustering at the end of P4, and there was clear consensus that action was needed. The early ideas behind the proposed clustering emerged from a back-of-napkin discussion at the WG/IG meeting in Washington DC, November 2014, with Beth Plale, Kathy Fontaine, Jay Pearlman, and Françoise Pearlman. Mark Parsons developed the notions further in front of the WG/IG group. TAB then formed a task force of TAB members Beth Plale, Peter Fox, Françoise Genova, Rainer Stotzka, and Peter Wittenburg, and Engagement IG Chair Inna Kouper, who met winter of 2014-15 to formulate the overall model.

TAB endorsed the overall model at its 21/22 January 2015 meeting and over the next several weeks the TAB liaisons provided input on the categorization of the WGs and IGs with which they interact as liaison.

Purpose: the purpose of the area clustering is to guide and inform. Specifically area clustering will have the following uses:

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. Help inform **WG/IG members** about other groups' activities.
5. Help **TAB** guide and assist existing and new groups.
6. Help **TAB** and WGs/IGs themselves identify gaps and overlaps in WG/IG activity.

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

II. Area Clustering

The Area Clustering is mapped into a two-dimensional space where it is made up of 4 quadrants. Each Working Group and Interest Group, as determined by their projected output products, occupies a point in a single quadrant of this space.

The two dimensions are: *solution dimension* (Y-axis) and a *beneficiary dimension* (X-axis). The ***solution dimension*** is a spectrum from technical to social, where a solution can manifest itself most strongly as software or infrastructure (technical), or as policy, governance, educational, or community building (social). The ***beneficiary dimension*** is a spectrum from data providers to data consumers, where the primary beneficiary is the data provider (or act of data provisioning) on one end, or the data consumer on another end. In many cases, both data provider and consumer benefit, in which case there may be a primary beneficiary, or an activity may sit in the middle.

Each quadrant is defined as follows:

Q1: Social/educationally oriented activity that benefits data consumer: products emerging from WG or IG in this quadrant are solutions to data sharing that benefit the data consumer more than the data provider, and manifests itself most strongly as new policy, governance, educational, or community building. Common terms include education, engagement, bridging, community

Q2: Technically-oriented solutions that benefit data consumer: products emerging from WG or IG in this quadrant are solutions to data sharing that benefit the data consumer more than data provider, and manifests itself most strongly as new approaches to data interoperability, harmonization, integration, or metadata.

Q3: Technical solutions that aid in data provisioning: products emerging from WG or IG in this quadrant are solutions to data sharing that benefit the data provider more than data consumer, and manifest themselves primarily technically through new software or infrastructure. Common terms include repository, fabric, analytics, identity, management.

Q4: Policy oriented solutions that aid in data provisioning: products emerging from WG or IG in this quadrant are solutions to data sharing that benefit the data provider more than the data consumer, and manifest themselves primarily through socially-oriented solutions (policy, governance, legal). Common terms include governance, certification, cost recovery, legal.

The four quadrants are shown in Figure 1 labeled by common terms appearing in the names of the WG and IGs that occupy the quadrant.

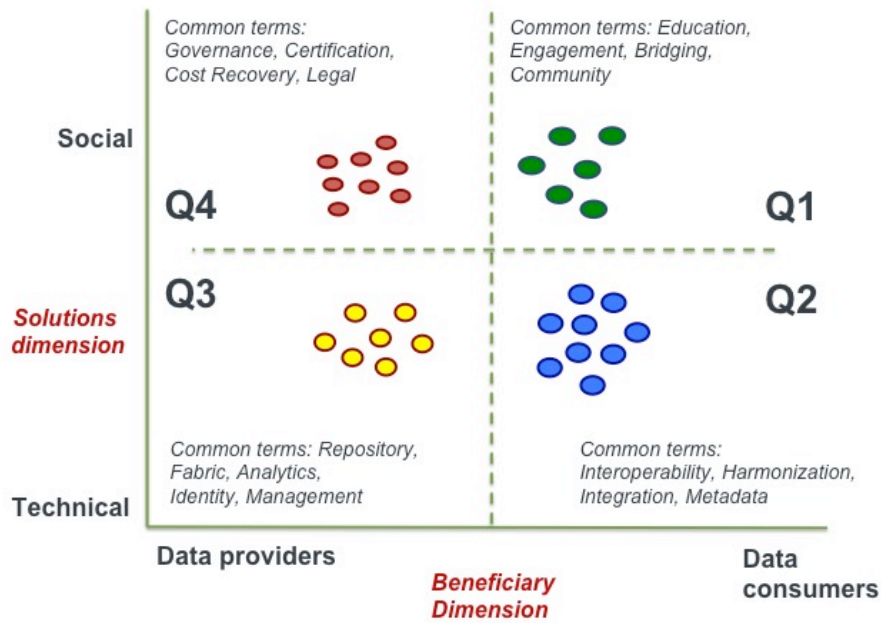


Figure 1. Solution-beneficiary clustering of WG/IG anticipated output products into quadrants

III. Positioning Working Groups and Interest Group

Each active Working Group is “binned” into one of the four quadrants as shown in Figure 2. Both active WGs and IGs are listed in their initial binning in Table 1. Common terms are underlined.

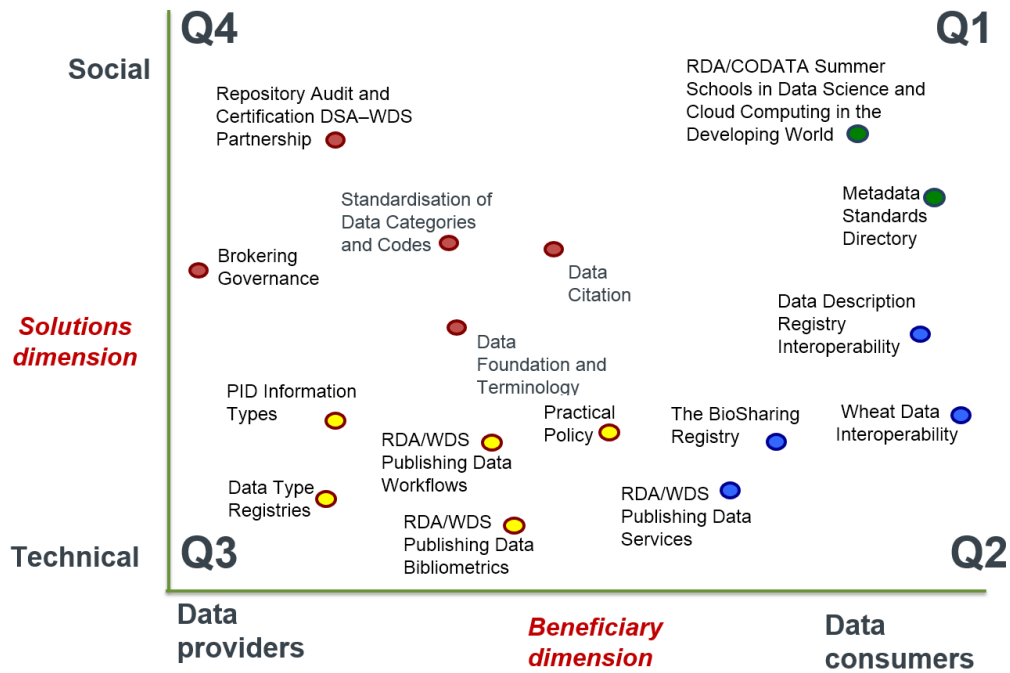






Figure 2. Working Groups Areas: note that the “binning” of WGs into areas is ongoing

IV. Tagging – Way to Further Describe

As a way for WGs and IGs to further describe their activities and outputs, we introduce unconstrained set of terms by which a WG/IG self-identifies. These terms can further categorize groups and aid navigation. A preliminary list of tags is shown below; 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) by Area (not final “binning”)

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

Appendix A. Supporting Documentation

Appendix A gives background of the different approaches that TAB considered in arriving at the proposed area clustering model.

Material in the Appendix is for background reading only

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.

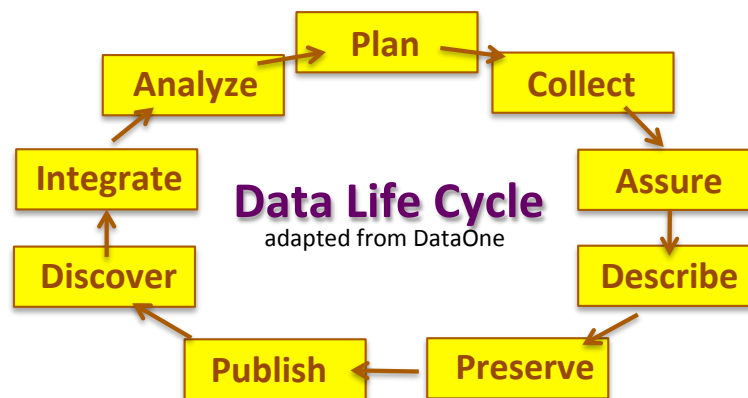


Figure 3. Data Lifecycle.

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

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

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?
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.

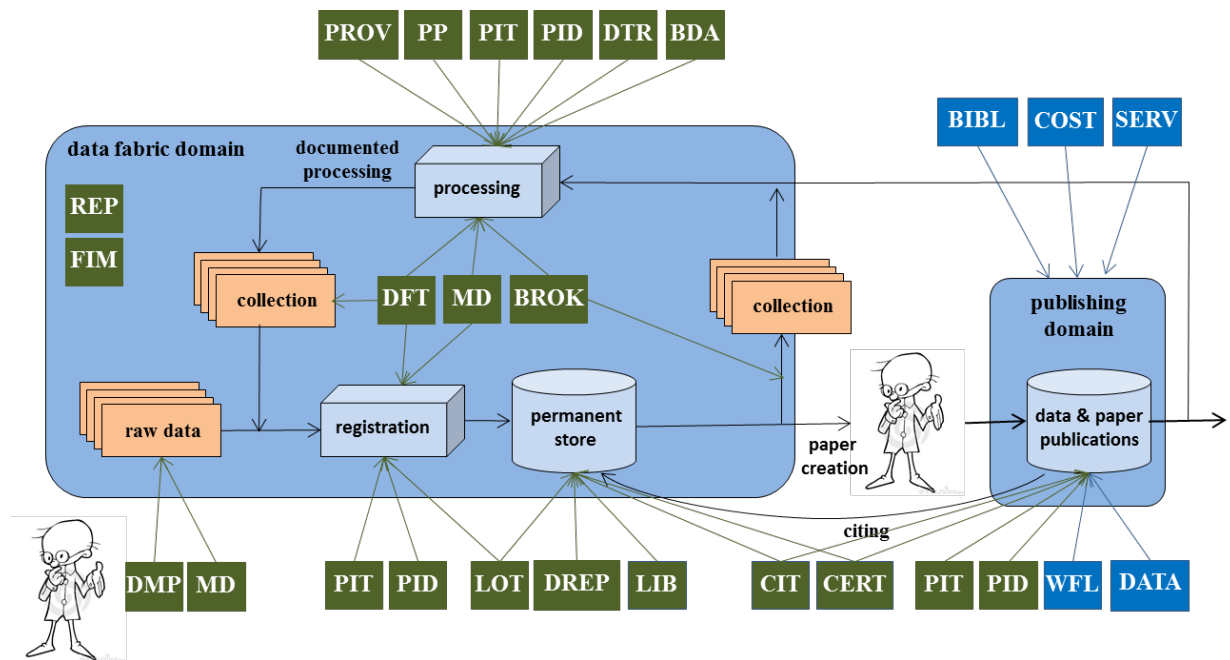


Figure 4. Functional phases of data; the following abbreviations are used in the diagram: **Brok**: Brokering WG & IG, **CIT**: Data Citation, **DFT**: Data Foundation & terminology, **DTR**: Data Type Registries, **MD**: Metadata WG & IGs, **PIT**: PID Information Types, **PP**: Practical Policies, **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	IG		PIDs	B
				WG	Metadata Standards Directory	A/D/E
				Data Description Registry Interoperability	A	
			IG		Research Data Provenance	A/D/E
					Data in Context	A/D/E
					Metadata	A/D/E
		registry	WG	Data Type Registries	D	
		workflow/processing	WG	Practical Policy	E	
			IG		Big Data Analytics	E
					Long tail of research data	E
	non-technical	publishing/citation	IG		Brokering	I
					Federated Identity Management	I/C
			WG	Preservation e-Infrastructure	G/H	
		quality	IG	Data Citation	A	
			IG	Publishing Data	A	
	legal	IG	Certification of Digital Repositories	G/H		
community	IG		Legal Interoperability	C		
			Community Capability Model	X		
			Development of cloud computing capacity and education for developing world	X		
		Engagement Group	X			
discipline-agriculture		WG	Wheat Data Interoperability	X		

specific groups			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