The Metadata Groups

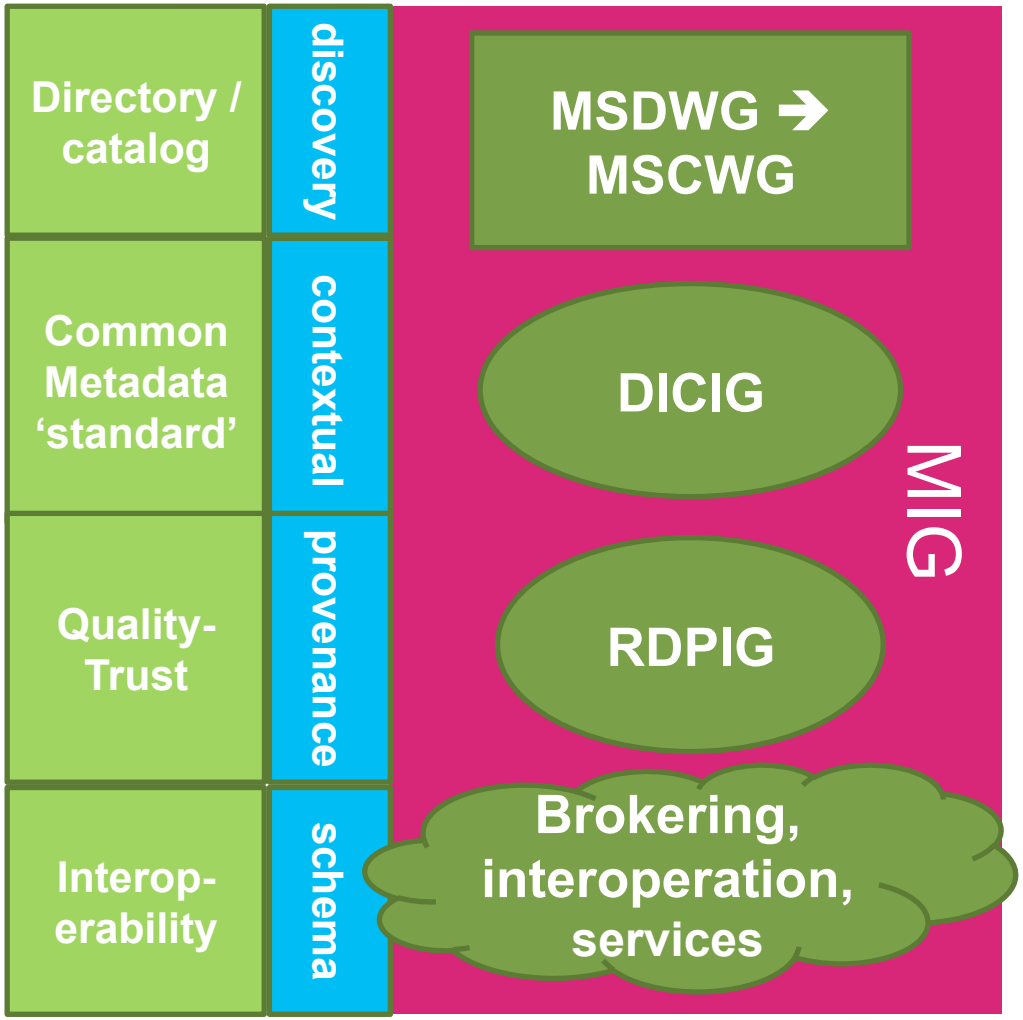
Keith G Jeffery

research data sharing without barriers

rd-alliance.org

Positioning

- Raise profile of metadata
  - Data first
  - Also software, resources, users
- Achieve outputs/ outcomes
- Present Plan
  - Involves other metadata groups
  - and most RDA groups



But actually……

Infrastructure Groups

Data Fabric
DFT
PID
Citation
Repositories
……..

Metadata Groups

Domain Groups

Biodiversity
Marine
Biosharing
Wheat
Neutrons
……..

Metadata Used for

- Discovery
  - Finding data, software, people (users), resources (computers, storage, instrumentation);
  - Relevance to purpose (initial assessment)
- Contextualisation
  - Appropriate for the purpose (assessment)
  - Quality (syntax, semantics – including multilinguality)
  - Recording activity (who did what, when, where to which)
- Access: Connecting data to Software / Environment / resources
  - Schema level (syntax, semantics – including multiliguality)

Metadata / other groups

- Plan
  - Involves not only metadata groups but all RDA
  - Evolved since P2 Washington DC
- Use cases into repository (DICIG)
- Standards into MSDWG directory/catalog (MSDWG ➔ MSCWG)
- Analyse for commonalities and differences (MIG)
- Propose canonical metadata 'packages' for 'purposes' (MIG) consisting of 'elements'
- Validation of 'packages' (domain groups)
- Provision of convertors
  - This is a resource problem!
- Move to standardisation of 'packages' (RDA)

Metadata Principles

Keith G Jeffery

research data sharing without barriers

rd-alliance.org

METADATA Principles
Created and endorsed by the RDA Metadata Groups

- The only difference between metadata and data is mode of use
- Metadata is not just for data, it is also for users, software services, computing resources
- Metadata is not just for description and discovery; it is also for contextualisation (relevance, quality, restrictions (rights, costs)) and for coupling users, software and computing resources to data (to provide a VRE)
- Metadata must be machine-understandable as well as human understandable for autonomicity (formalism)
- Management (meta)data is also relevant (research proposal, funding, project information, research outputs, outcomes, impact…)

RESEARCH DATA ALLIANCE

Metadata Elements for Packages for Purposes

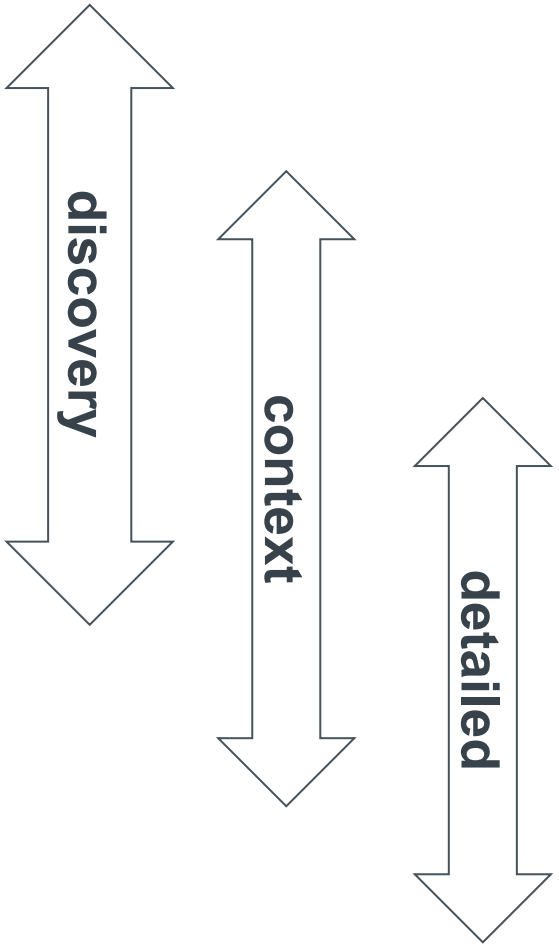
Keith G Jeffery

research data sharing without barriers

rd-alliance.org

Open Data: Purposing the elements

- Unique Identifier (for later use including citation)
- Location (URL)
- Description
- Keywords (terms)
- Temporal coordinates
- Spatial 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



But Note

- Many (most) of the elements are not simple single valued
  - e.g. Originator (organisation(s) / person(s))
- Many are multilingual
- So we need structured metadata (not flat metadata)
  - Base entities (exist in real world)
  - Linking entities (relationships between base entities) with role and temporal duration

Analysis of Use Cases (and to some extent, standards)

Keith G Jeffery, Rebecca Koskela

research data sharing without barriers

rd-alliance.org

Method

- Use Case Templates
  - Aligned elements and added additional elements
  - Referenced to a few standards
  - ➔ sheet 1
- Extracted common and required elements
  - ➔ sheet 2
- Considered structure utilising modern semantic linking
  - ➔ sheet 3
- Reduced to minimum that covers requirements of use cases
  - ➔ sheet 4

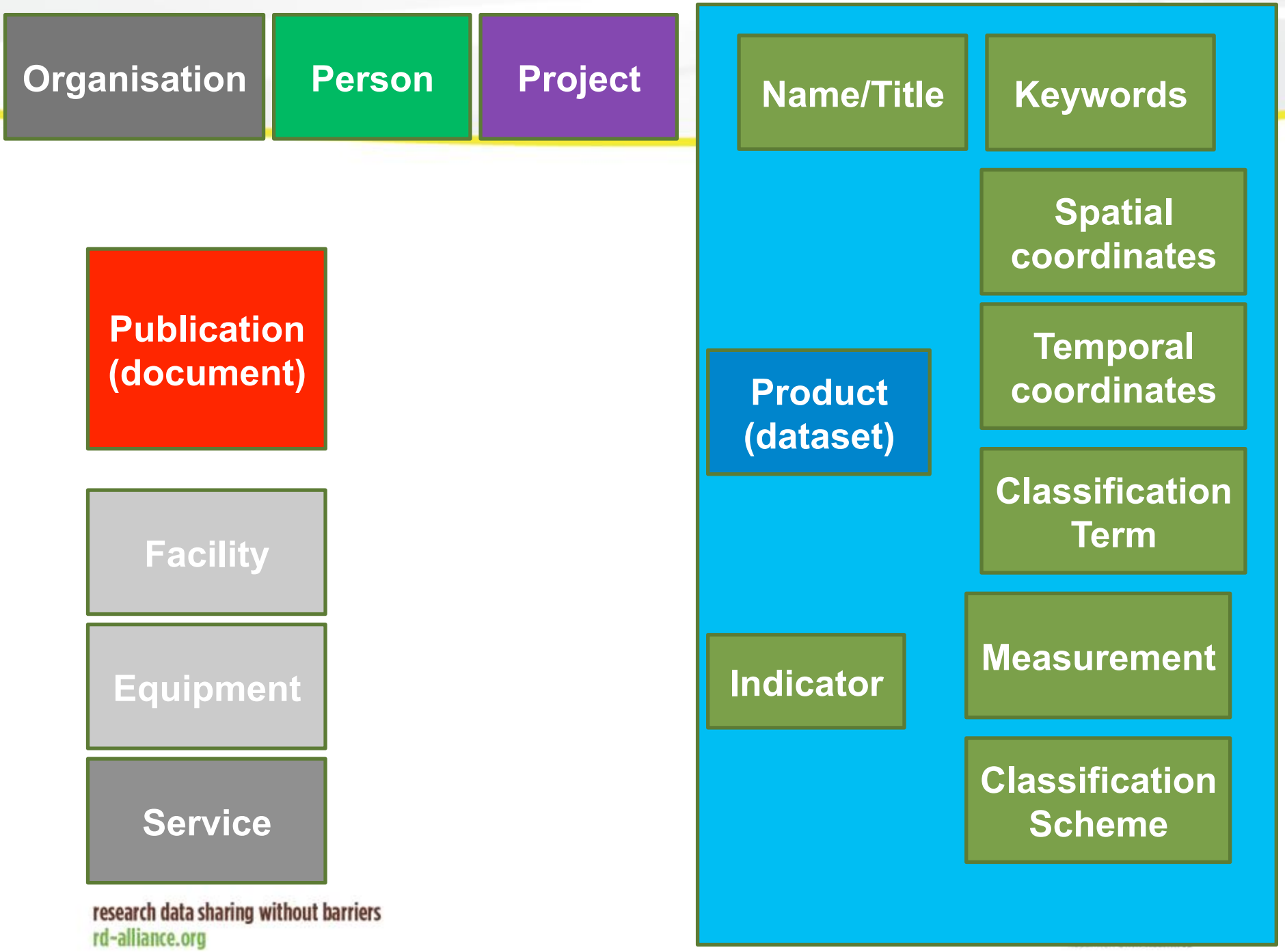
Spreadsheets

Name	1: Register a dataset rescued from legacy literature	2: PDB	3: Bore-hole	4: Library: Putting a dataset on a dedicated server	5: Event log processing of Hospital business processing event logs	6: intervention study survey	7: Agricul data or CIARD
Unique Identifier (for later use including citation)	Y	Y	y		Y	y	
Location (URL)	Y		n		Y	y	Y
Description	Y	Y	n		Y	y	
Keywords							
Temporal							
Geospatia							
Originator							
Project							
Facility /							
Quality							
Availability							

PROPOSED ELEMENTS		VIEWPOINT DATASET OUTWARDS
Name		
Unique Identifier (for later use including citation)	1:n	require federated IDs linked to the primary ID with roles; including e.g. DoI
Location (URL)	"1:1	
Title	1:n	require multilinguality
Description	1:n	require multilinguality
Keywords (terms)	1:n:m	multiple keywords and require multilinguality
Tempo		
Geosp		
Origin		
Projec		
Facilit		

"1:n	N:m		Notes		Also related to
Entity (language variants)	Entity	LinkingEntity(ies)			
Name/Title					
Description					
	Keywords (terms)	Dataset-Keyword			
		(all) Temporal coordinates	start/end	yymmddhhmmss	in linking entities

REDUCTION TO PROPOSAL			
ProductID		(dataset, could also be software or other product; use of 'product' also allows collections (structures); type of product defined by classification)	
		(below elements represented NOT as attributes but as relationshps to preserve referential and functional integrity)	
		(linking relationships have role and also date/time start, date/time end)	
	1:n	Name/Title	(multilinguality)
	1:n:m	Keywords	(multilinguality)
	n:m	Geospatial coordinates	(more work needed, complex structure, need coordinate system, accuracy, precision...; for earth ISO19115/ISPIRE, for space?
	n:m	Temporal coordinates	(more work needed; temporal coordinates can be handled by relationships: difference between event (when collected) and te
	n:m	Organisation	
	n:m	Person	


Organisation
Person
Project
Publication (document)
Facility
Equipment
Service
Name/Title
Keywords
Spatial coordinates
Temporal coordinates
Product (dataset)
Classification Term
Indicator
Measurement
Classification Scheme

Organisation

Person

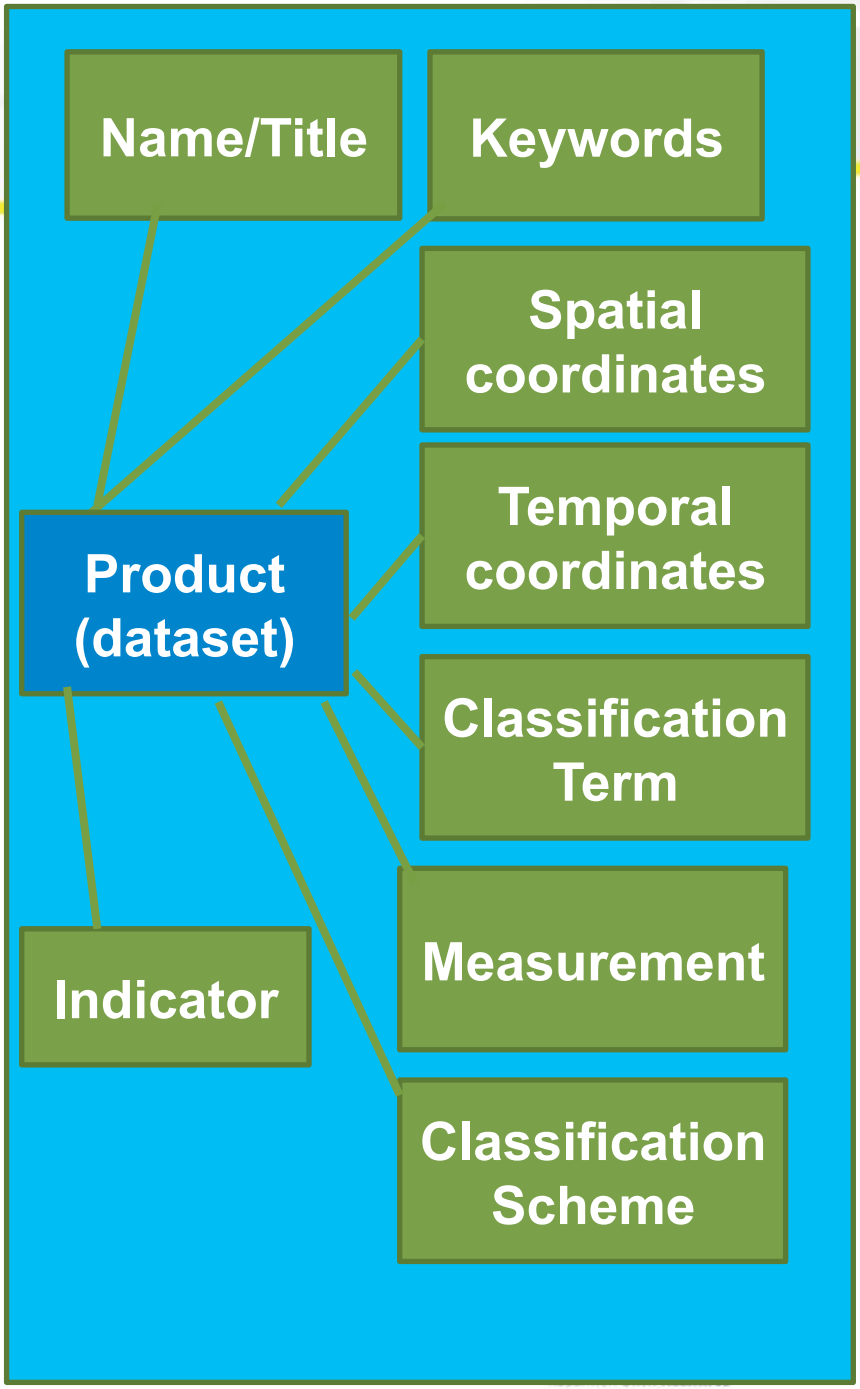
Project

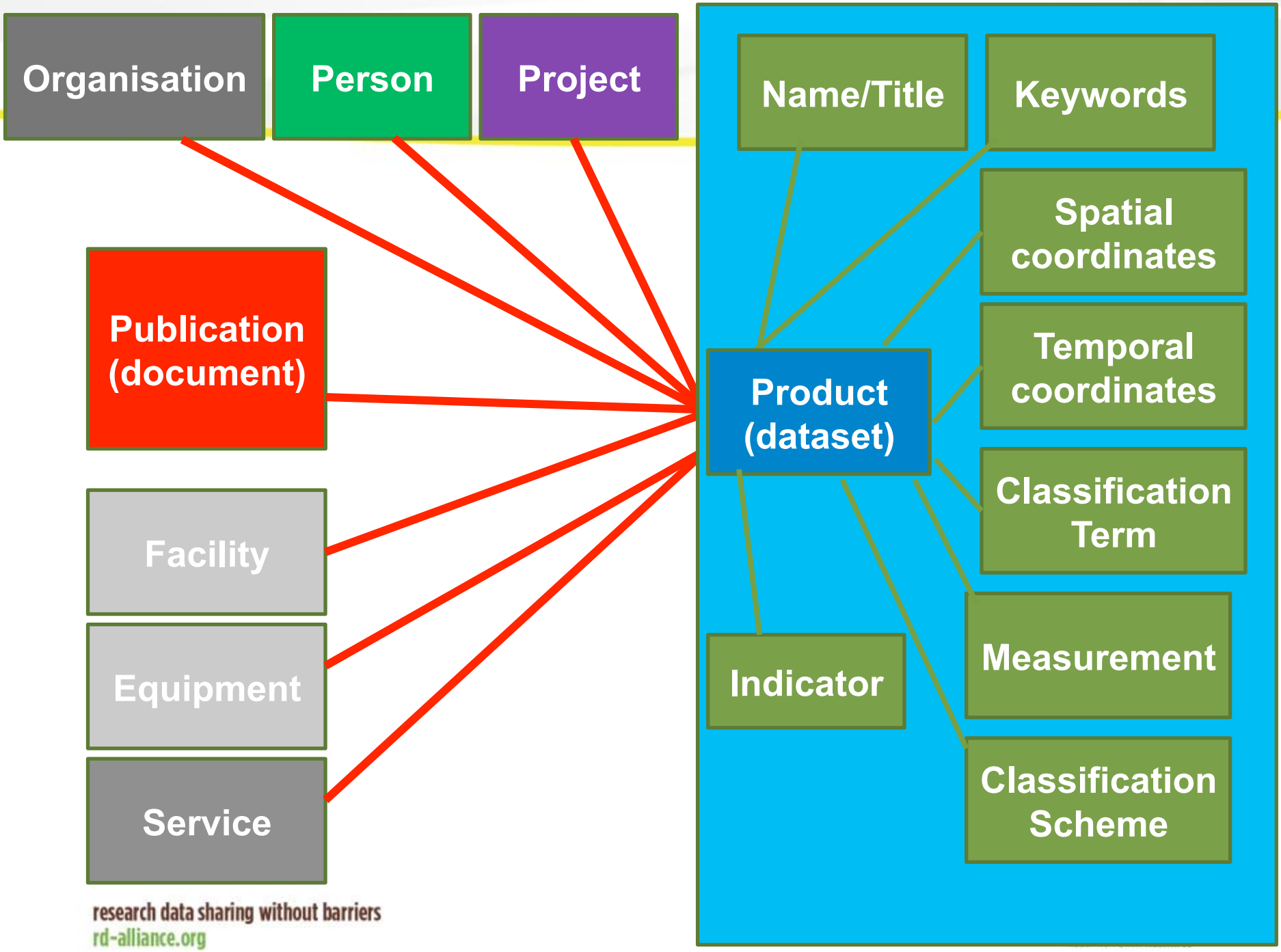
Publication (document)

Facility

Equipment

Service




Organisation
Person
Project
Publication (document)
Facility
Equipment
Service
Name/Title
Keywords
Spatial coordinates
Temporal coordinates
Classification Term
Product (dataset)
Indicator
Measurement
Classification Scheme

Linking

- The lines drawn connect two entities indicating a relationship;
- There may be many linking relationships between the same two entities;
- Each has a role e.g. Person-OWNER-Dataset;
  - The roles are defined in some kind of ontology
- Each has a start and end date/time
  - This allows for versioning and provenance tracking

Discussion

- Do we need licence as a separate entity (e.g. CC-BY-NC) or is the entity publication/document (with its name including licence kind) sufficient?
- Do we need language as a separate attribute describing the language of the dataset content or is language a classification term associated to the dataset? What happens if the dataset has attribute values in>1 language.
  - Note: multiple versions in different languages – presumably each of these is identified uniquely?
- Do we need subject (what the dataset is recording – ideally from a defined list of terms e.g. blood samples) as an entity or attribute or as a classification term associated to the dataset?