The Metadata Groups

Keith G Jeffery
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......
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
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...)

METADATA Principles
Created and endorsed by the RDA Metadata Groups
Metadata Elements for Packages for Purposes
Keith G Jeffery
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
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
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

### PROPOSED ELEMENTS

<table>
<thead>
<tr>
<th>Name</th>
<th>Unique Identifier (for later use including citation)</th>
<th>Location (URL)</th>
<th>Description</th>
<th>Keywords (terms)</th>
<th>Temporal coordinates</th>
<th>Geospatial coordinates</th>
<th>Originator</th>
<th>Project</th>
<th>Facility</th>
<th>Notes</th>
<th>Also related to</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### VIEWPOINT DATASET OUTWARDS

1:1

<table>
<thead>
<tr>
<th>Name</th>
<th>Unique Identifier (for later use including citation)</th>
<th>Location (URL)</th>
<th>Title</th>
<th>Description</th>
<th>Keywords (terms)</th>
<th>Temporal coordinates</th>
<th>Geospatial coordinates</th>
<th>Originator</th>
<th>Project</th>
<th>Facility</th>
<th>Notes</th>
<th>Also related to</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### REDUCTION TO PROPOSAL

1:n

<table>
<thead>
<tr>
<th>Product-ID</th>
<th>(dataset, could also be software or other product; use of 'product' also allows collections (structures); type of product defined by classification)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(below elements represented NOT as attributes but as relationships to preserve referential and functional integrity)</td>
</tr>
<tr>
<td></td>
<td>(linking relationships have role and also date/time start, date/time end)</td>
</tr>
</tbody>
</table>

- 1:n Name/Title (multilinguality)
- 1:n Keywords (multilinguality)
- m:n Geospatial coordinates (more work needed, complex structure, need coordinate system, accuracy, precision... for earth ISO19115/SPIRE, for space?)
- m:n Temporal coordinates (more work needed; temporal coordinates can be handled by relationships; difference between event (when collected) and te
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?