Metadata Management
Concepts and Principles

Jon Johnson
12 September 2019
CLOSER, UCL Institute of Education
Connecting up research
Relationships

• Metadata provides information enabling it to make sense of:
  • **data** (e.g. documents, images, datasets),
  • **concepts** (e.g. controlled vocabularies, classification schemes),
  • **real-world entities** (e.g. people, organisations, places)
  • **processes** (e.g. data collections, archiving, computation)

• The relationship between these information items allows us to manage them

• Managing metadata could be considered to be managing the relationships between different types of content
Concepts

• Concepts are defined in ISO/IEC 11179
  • Unit of knowledge created by a unique combination of characteristics
  • Concepts are not necessarily bound to particular languages. They are, however, influenced by the social or cultural background which often leads to different categorizations
  • A concept is independent of its representation
  • Not Keywords
  • A concept may have sub-concepts

Examples are: height, weight, country, sex
Concepts – Example

:MaritalStatus rdf:type skos:ConceptScheme.
:Married rdf:type skos:Concept;
    skos:inScheme :MaritalStatus;
    skos:related :Single .
:Single rdf:type skos:Concept;
    skos:inScheme :MaritalStatus;
    skos:narrower :NeverMarried .
:NeverMarried rdf:type skos:Concept;
    skos:inScheme :MaritalStatus;
:Widowed rdf:type skos:Concept;
    skos:inScheme :MaritalStatus;
    skos:broader :Single .
:Divorced rdf:type skos:Concept;
    skos:inScheme :MaritalStatus;
    skos:broader :Single .
Concepts - Controlled Vocabularies

Controlled vocabularies are used in descriptive metadata fields to support consistent, accurate, and quick indexing and retrieval of digital asset content. It has a specific definition associated with a particular value*. 

<table>
<thead>
<tr>
<th>Value</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>Two or more people related by blood, marriage (including step-relations), adoption or fostering and who may or may not live together. For example, used when researching the extent to which people provide support and assistance for their relatives.</td>
</tr>
<tr>
<td>Household</td>
<td>A person or a group of persons who share the same dwelling unit and common living arrangements. These common living arrangements may include pooling some, or all, of their income and wealth, and consuming certain types of goods and services collectively, mainly housing and food</td>
</tr>
</tbody>
</table>

Unit Type

• Unit type is a synonym of Object class defined in ISO/IEC 11179
  • A sub-type of concept
  • A class of object of interest
  • It is used to describe a class or group of Units based on a single characteristic, but with no specification of time and geography.
  • For example, the Unit Type of “Person” groups together a set of Units based on the characteristic that they are ‘Persons’.

Examples are: Person, Establishment, Household, State, Country, Dog, Automobile
Unit type – Example
Universe and Population

• A Universe is a set of entities defined by a more narrow specification than that of an underlying UnitType.

• A Population further narrows the specification to a specific time and geography.

• Universe sits in a hierarchy between UnitType and Population, with UnitType being most general and Population most specific.
Universe – Example
Category

• A category is concept whose role is to define and measure a characteristic. It has a reference to its underlying concept.

• Examples
  • Afghanistan, Trinidad and Tobago,
  • female, male
  • Cox’s Pippin, Golden Delicious
  • Sausage, Afghan
Categories as concepts

Concept

Are you?

Categories

If single, are you?

- Married
- Single
- Never married
- Widowed
- Divorced
Code Lists

- Category is the meaning associated with a Code – it is a specialized type of Concept.
- When the Code appears in a data set (as a variable value) or questionnaire (as a response to a question), it is referring to the meaning supplied by the Category.
- The Code is just the sign which refers to the Category – in and of itself, it has no meaning/definition. That is supplied by the Category.
- A code value must be unique within a code list
- Code lists may be flat or hierarchical
Code lists – Examples

ISO/IEC 5218 - Codes for the representation of human sexes

0 Not known
1 Male
2 Female
9 Not applicable

ISO 3166-1 Alpha-2 - Codes for the representation of names of countries and their subdivisions

AP - Afghanistan
TT - Trinidad and Tobago
VU - Vanuatu

ISO 3166-1 numeric - Codes for the representation of names of countries and their subdivisions

004 - Afghanistan
780 - Trinidad and Tobago
548 - Vanuatu

Nomenclature of Economic Activities
(European statistical classification of economic activities)

<table>
<thead>
<tr>
<th>Division</th>
<th>Group</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>01.1</td>
<td>Crop and animal production, hunting and related service activities</td>
</tr>
<tr>
<td></td>
<td>01.11</td>
<td>Growing of non-perennial crops</td>
</tr>
<tr>
<td></td>
<td>01.12</td>
<td>Growing of cereals (except rice), leguminous crops and oil seeds</td>
</tr>
<tr>
<td></td>
<td>01.13</td>
<td>Growing of vegetables and melons, roots and tubers</td>
</tr>
<tr>
<td></td>
<td>01.14</td>
<td>Growing of sugar cane</td>
</tr>
<tr>
<td></td>
<td>01.15</td>
<td>Growing of tobacco</td>
</tr>
<tr>
<td></td>
<td>01.16</td>
<td>Growing of fibre crops</td>
</tr>
<tr>
<td></td>
<td>01.19</td>
<td>Growing of other non-perennial crops</td>
</tr>
<tr>
<td></td>
<td>01.2</td>
<td>Growing of perennial crops</td>
</tr>
<tr>
<td></td>
<td>01.21</td>
<td>Growing of grapes</td>
</tr>
<tr>
<td></td>
<td>01.22</td>
<td>Growing of tropical and subtropical fruits</td>
</tr>
<tr>
<td></td>
<td>01.23</td>
<td>Growing of citrus fruits</td>
</tr>
</tbody>
</table>

How satisfied are you with our services?

Very Unsatisfied
Unsatisfied
Neutral
Satisfied
Very Satisfied
The combination of these foundational objects, allows us to refine other objects and can be repurposed to improve data management, discovery and analysis.
Identification and Versioning

- Managing metadata could be considered to be managing the relationships between different types of content
- This content may change over time
- It is advisable to decouple the content from how it is referenced
- Each element therefore needs an ID
- If you want to know what has changed then you must version
- If you wish to know why it has changed, you should have a rationale that is aligned with the version.
Identification and Versioning

• The details of this are laid out in ISO/IEC 11179
  • Each administered item shall have a unique data identifier within the register of a Registration Authority

• An ID should therefore be composed of
  • An agency : an identifier : a version
    e.g. urn:ddi:agency:identifier:version

• This allows you to share metadata with another agency, safe in the knowledge that there will not be a clash of identifiers
Benefits of identifying items

Allows:

• management of items separately from each other
• reuse of items
• tracking of provenance between items
• use of a specific version of an item
• tracking and management of changes to content over time
Decouple the content from the reference - Example

• A response domain can be maintained separately from the question referencing it.

V1, V2, V3 represent different versions of the item.

People in general should care more about the environment

<table>
<thead>
<tr>
<th>Strongly agree</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly disagree</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>
Reusing objects

- Reusing objects means you don’t have duplication
- But referencing objects means you have to take care when you change them, (you need to be sure you are not changing something which is used elsewhere)
- If you want to change something, you can ‘take a copy’ and provide a reference to the original object using ‘BasedOn’. This provides a provenance chain.
Reuse of items - Example

- A response domain can be reused *by reference* by many different question items

- **People in general should care more about the environment**
- **Climate change is caused by human activity alone**
- **There is nothing I can do to prevent global temperature rise**

- **Strongly agree**: 0 1 2 3 4 5 6 7
- **Strongly disagree**: 8 9 10
Track provenance between items - Example

- Two related code lists with education categories. The short list is based on the long list. The short list has a **based-on** reference to the long list.
- This allows to track the relationship between the two related code lists.

### ID1 V1

<table>
<thead>
<tr>
<th>Code</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Not completed ISCED level 1</td>
</tr>
<tr>
<td>113</td>
<td>ISCED 1, completed primary education</td>
</tr>
<tr>
<td>129</td>
<td>Vocational ISCED 2C &lt; 2 years, no access ISCED 3</td>
</tr>
<tr>
<td>212</td>
<td>General/pre-vocational ISCED 2A/2B, access ISCED 3 vocational</td>
</tr>
<tr>
<td>213</td>
<td>General ISCED 2A, access ISCED 3A general/all 3</td>
</tr>
<tr>
<td>221</td>
<td>Vocational ISCED 2C &gt;= 2 years, no access ISCED 3</td>
</tr>
<tr>
<td>222</td>
<td>Vocational ISCED 2A/2B, access ISCED 3 vocational</td>
</tr>
<tr>
<td>223</td>
<td>Vocational ISCED 2, access ISCED 3 general/all</td>
</tr>
<tr>
<td>229</td>
<td>Vocational ISCED 3C &lt; 2 years, no access ISCED 5</td>
</tr>
<tr>
<td>311</td>
<td>General ISCED 3 &gt;=2 years, no access ISCED 5</td>
</tr>
<tr>
<td>312</td>
<td>General ISCED 3A/3B, access ISCED 5B/lower tier 5A</td>
</tr>
<tr>
<td>313</td>
<td>General ISCED 3A, access upper tier ISCED 5A/5B/lower tier 5A</td>
</tr>
<tr>
<td>321</td>
<td>Vocational ISCED 3C &gt;=2 years, no access ISCED 5</td>
</tr>
<tr>
<td>322</td>
<td>Vocational ISCED 3A, access ISCED 5B/lower tier 5A</td>
</tr>
<tr>
<td>323</td>
<td>Vocational ISCED 3A, access upper tier ISCED 5A/5B/lower tier 5A</td>
</tr>
<tr>
<td>412</td>
<td>General ISCED 4A/4B, access ISCED 5B/lower tier 5A</td>
</tr>
<tr>
<td>413</td>
<td>General ISCED 4A, access upper tier ISCED 5A/5B/lower tier 5A</td>
</tr>
<tr>
<td>421</td>
<td>ISCED 4 programmes without access ISCED 5</td>
</tr>
<tr>
<td>422</td>
<td>Vocational ISCED 4A/4B, access ISCED 5B/lower tier 5A</td>
</tr>
<tr>
<td>423</td>
<td>Vocational ISCED 4A, access upper tier ISCED 5A/5B/lower tier 5A</td>
</tr>
<tr>
<td>510</td>
<td>ISCED 5A short, intermediate/academic/general tertiary below bachelor</td>
</tr>
<tr>
<td>520</td>
<td>ISCED 5B short, advanced vocational education</td>
</tr>
<tr>
<td>610</td>
<td>ISCED 5A medium, bachelor/equivalent from lower tier tertiary</td>
</tr>
<tr>
<td>620</td>
<td>ISCED 5A medium, bachelor/equivalent from upper/single tier tertiary</td>
</tr>
<tr>
<td>710</td>
<td>ISCED 5A long, master/equivalent from upper/single tier tertiary</td>
</tr>
<tr>
<td>720</td>
<td>ISCED 5A long, master/equivalent from upper/single tier tertiary</td>
</tr>
<tr>
<td>800</td>
<td>ISCED 6 Doctoral degree</td>
</tr>
</tbody>
</table>

### ID2 V1

<table>
<thead>
<tr>
<th>Code</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ES-ISCED I, less than lower secondary</td>
</tr>
<tr>
<td>2</td>
<td>ES-ISCED II, lower secondary</td>
</tr>
<tr>
<td>3</td>
<td>ES-ISCED IIIb, lower tier upper secondary</td>
</tr>
<tr>
<td>4</td>
<td>ES-ISCED IIIa, upper tier upper secondary</td>
</tr>
<tr>
<td>5</td>
<td>ES-ISCED IV, advanced vocational, sub-degree</td>
</tr>
<tr>
<td>6</td>
<td>ES-ISCED V1, lower tertiary education, BA level</td>
</tr>
<tr>
<td>7</td>
<td>ES-ISCED V2, higher tertiary education, &gt;= MA level</td>
</tr>
</tbody>
</table>

International Standard Classification of Education
Summary of versioning and identification

• Metadata content should be decoupled from how it is referenced to allow:
  • Change management
  • Re-use
  • Provenance

This puts your metadata on a firm footing for the current (and future) generation of semantic technologies