

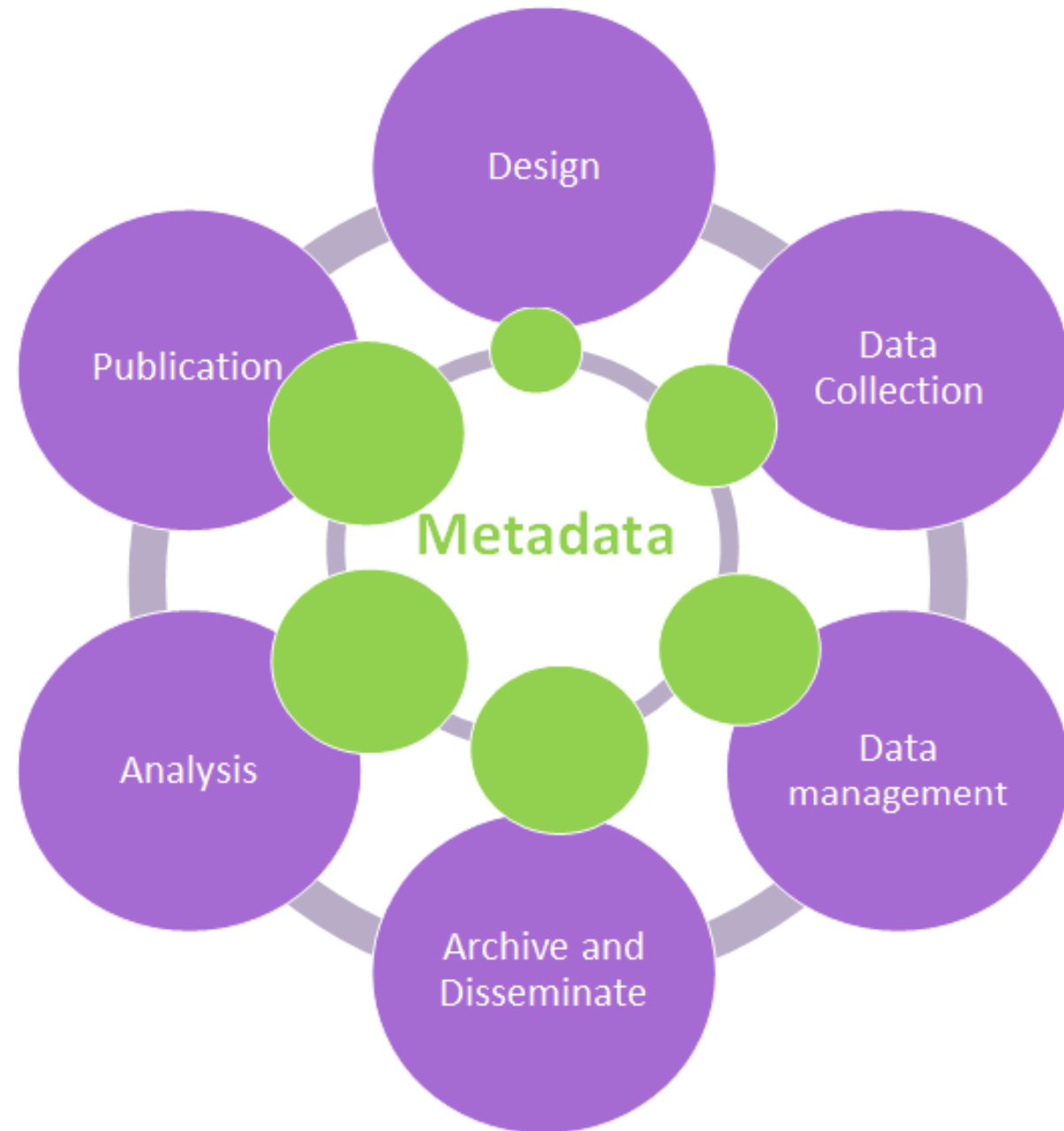
Metadata Management Concepts and Principles

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



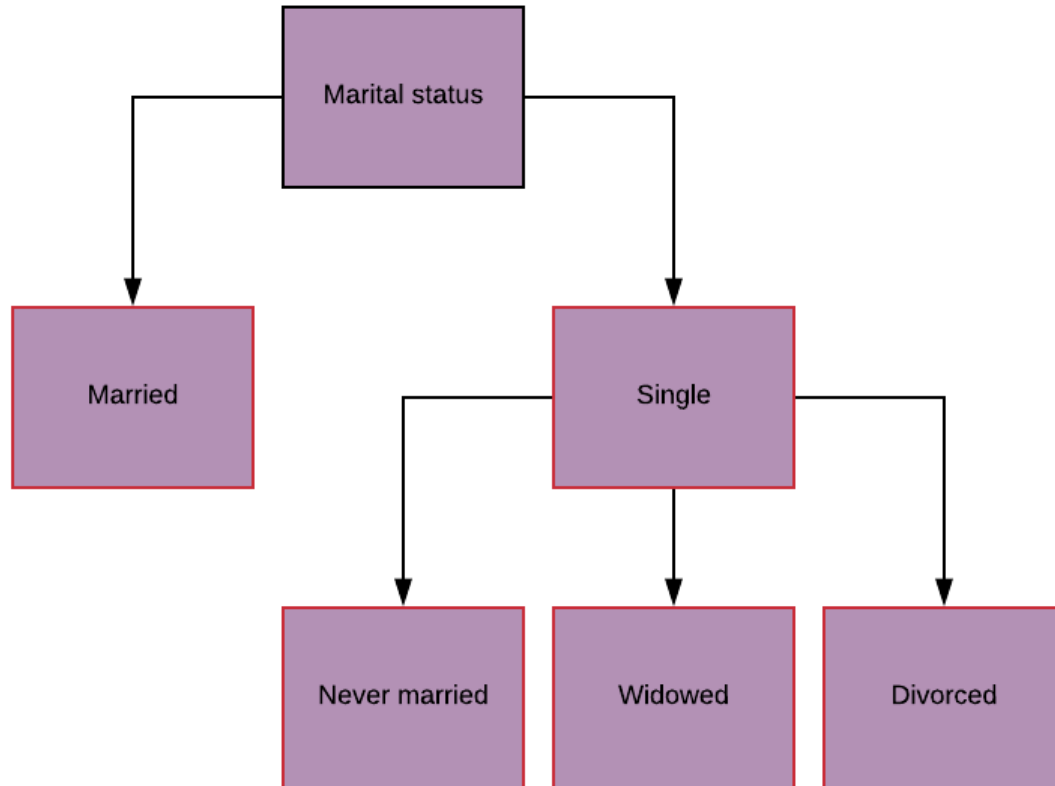
Concepts

- 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*.

Value	Definition
Family	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.
Household	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

* Hedden, H (2007) Taxonomies and controlled vocabularies best practices for metadata. (<https://link.springer.com/article/10.1057/dam.2010.29>)



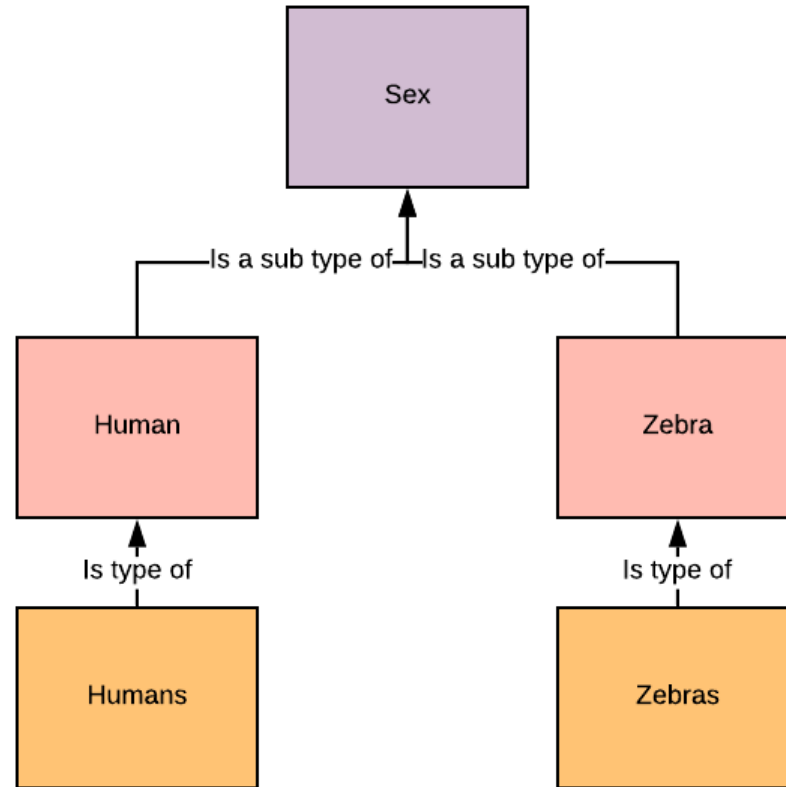
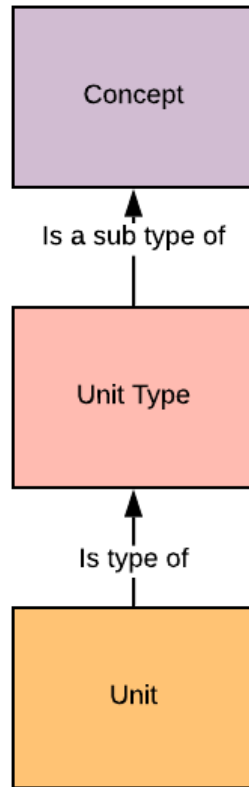
Unit Type

- 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



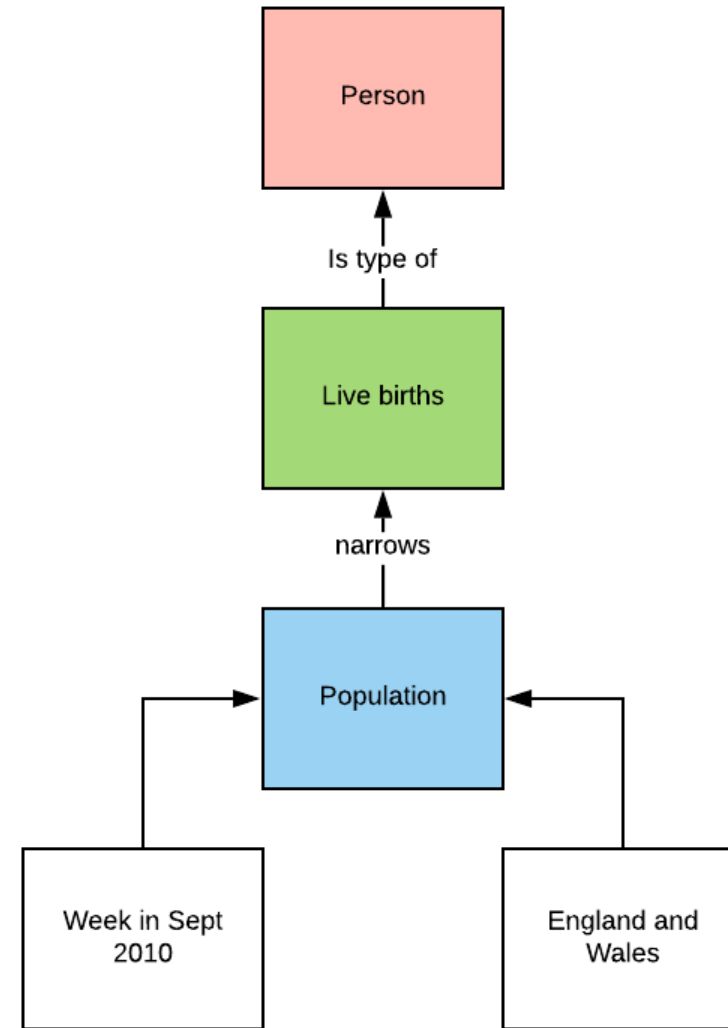
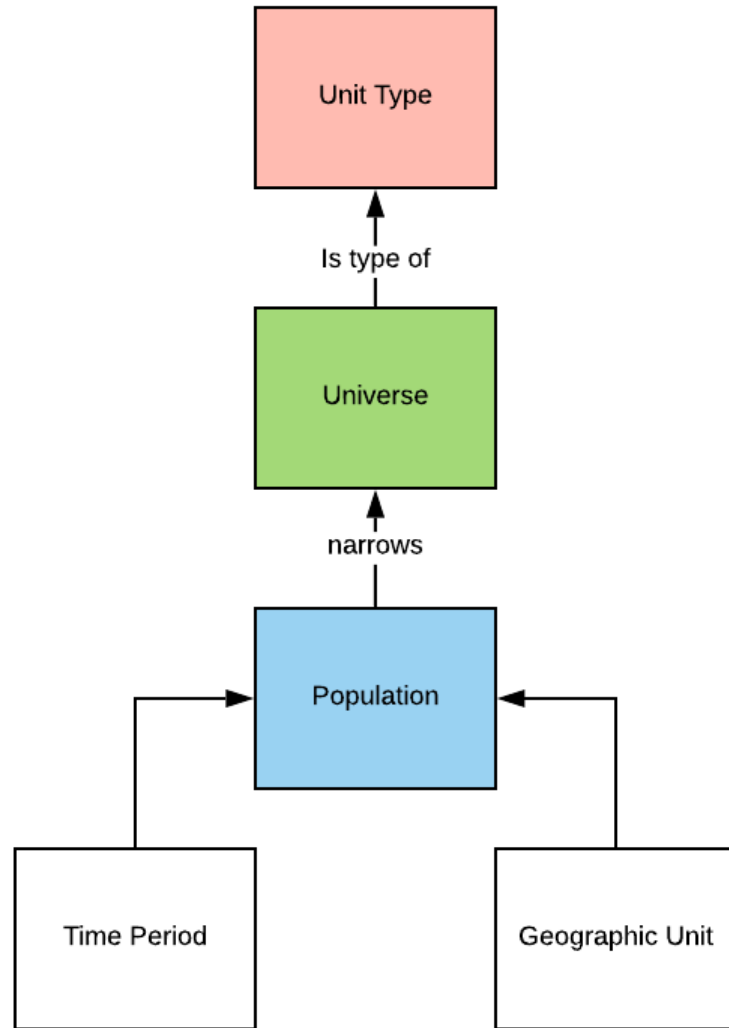


Universe and Population

- 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



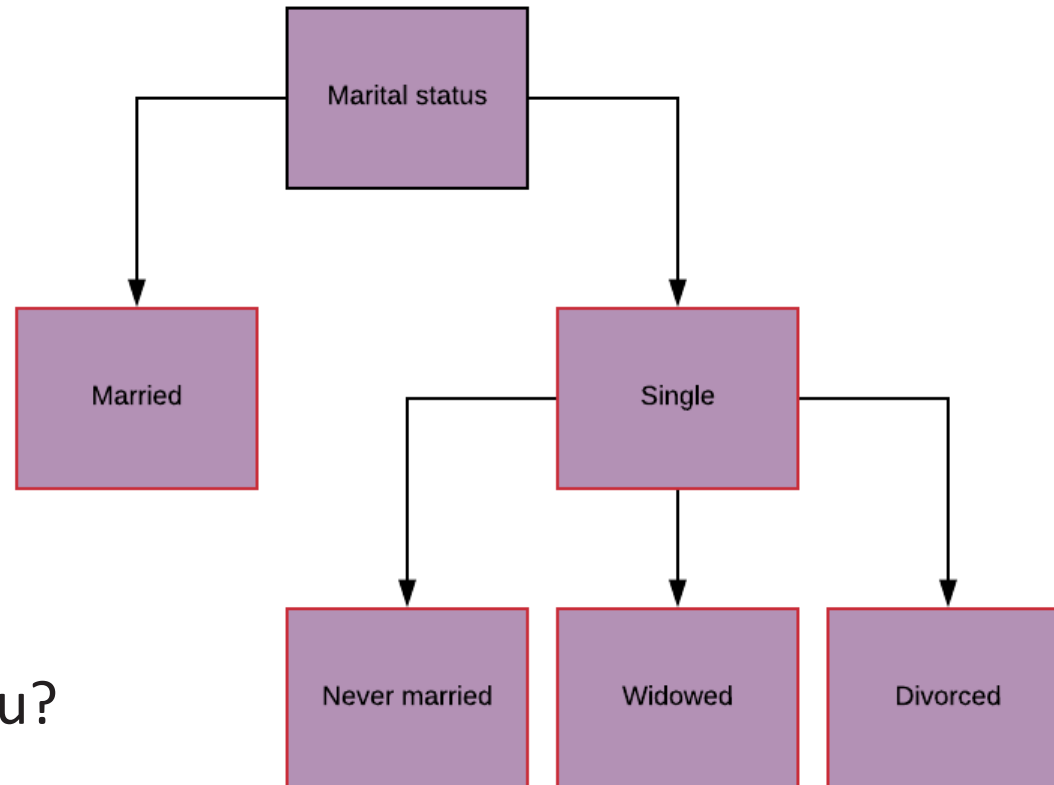
Categories as concepts

Concept

Categories

Are you?

If single, are you?





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)

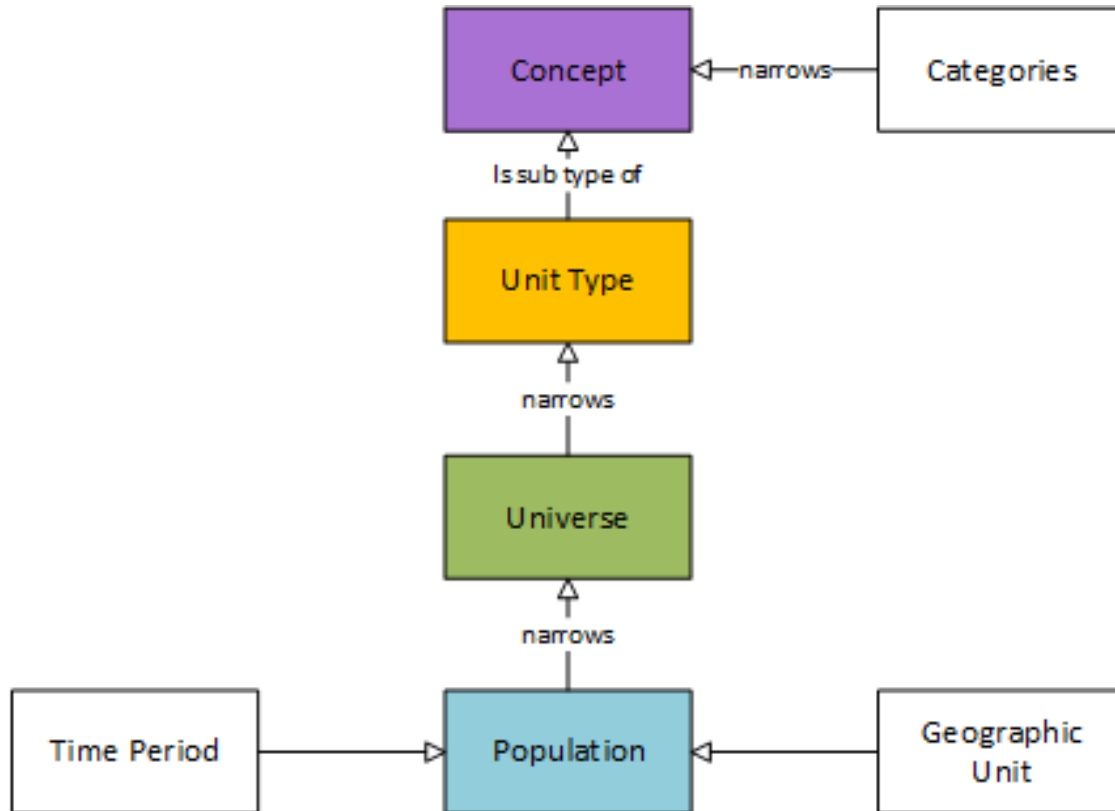
Division	Group	Class	
SECTION A — AGRICULTURE, FORESTRY AND FISHING			
01	01.1		Crop and animal production, hunting and related service activities
			Growing of non-perennial crops
		01.11	Growing of cereals (except rice), leguminous crops and oil seeds
		01.12	Growing of rice
		01.13	Growing of vegetables and melons, roots and tubers
		01.14	Growing of sugar cane
		01.15	Growing of tobacco
		01.16	Growing of fibre crops
		01.19	Growing of other non-perennial crops
		01.2	Growing of perennial crops
		01.21	Growing of grapes
		01.22	Growing of tropical and subtropical fruits
		01.23	Growing of citrus fruits

How satisfied are you with our services?

				
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Very Unsatisfied	Unsatisfied	Neutral	Satisfied	Very Satisfied



Summary



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.





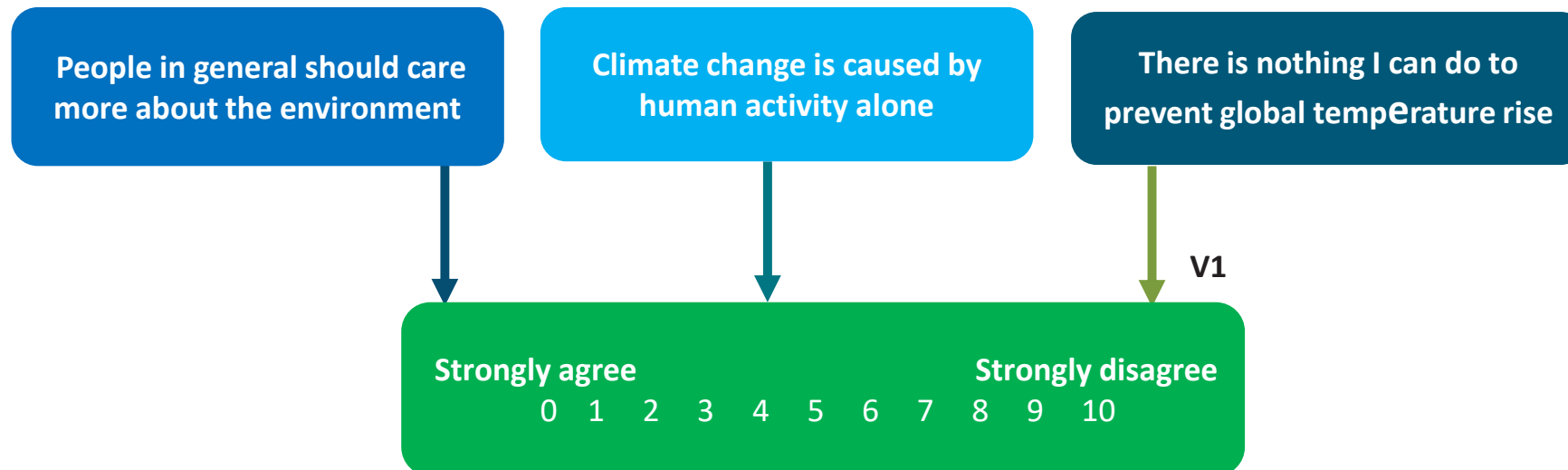
Reusing objects

- 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





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.

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

BasedOn

ID2 V1

code	category
1	ES-ISCED I, less than lower secondary
2	ES-ISCED II, lower secondary
3	ES-ISCED IIIb, lower tier upper secondary
4	ES-ISCED IIIa, upper tier upper secondary
5	ES-ISCED IV, advanced vocational, sub-degree
6	ES-ISCED V1, lower tertiary education, BA level
7	ES-ISCED V2, higher tertiary education, >= MA level

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