Project N°: 262608

# ACRONYM: Data without Boundaries

## DELIVERABLE D8.1

*(OS Object Model)*

## WORK PACKAGE 8

*(Improving Resource Discovery for OS Data)*

| | | |
|---|---|---|
| REPORTING PERIOD: | From: Month 1 | To: Month 18 |
| PROJECT START DATE: | 1st May 2011 | DURATION: 48 Months |
| DATE OF ISSUE OF DELIVERABLE: | 24.04.2012 | |
| DOCUMENT PREPARED BY: | 16, 14, 27, 19 | Metadata Technology Ltd, Centro De Investigaciones Sociologicas, Centrul National De Pregatire In Statistica, Koninklijke Nederlandse Akademie Van Wetenschappen - KNAW |

Combination of CP & CSA project funded by the European Community
Under the programme "FP7 - SP4 Capacities"
Priority 1.1.3: European Social Science Data Archives and remote access to Official Statistics

## CONTENTS

## FIGURES

## I. Introduction

This model defines the types of first-order metadata objects we will need to support search and discovery across official statistical sources of microdata, as well as those from data archives and research centres. It covers both aggregate and micro-data sets, and attempts to capture the key aspects of the SDMX[1] model, as that will be the typical model for aggregate official statistics, and also emphasizes similarities with the SDMX-based Data Cube vocabulary when using RDF[2] expressions of the metadata.

On the microdata side, the key objects for search and discovery have been modelled, with an attempt to show how these objects might relate. Note that this is a very simplified view, which describes nothing below the logical level – if the researcher wishes to use the data after he/she has discovered it, which the researcher would need additional physical descriptions of the file, presumably supplied by the disseminator of the data.

Note that this model may contain some fields that are inserted to support the use cases of other work packages (notably work package 5). Once all work package (WP) 8 requirements are identified, the model should be examined to determine what is not required for implementation of the portal in WP12. This will result in the more detailed implementation model.

## II. The Style of UML Used in The Model

Top-level objects are depicted as classes, with three different types of relationships. Although there is often a distinction made between composition relationships and aggregation relationships, the style of UML[3] used here is similar to that found in the SDMX information model: aggregations are not used – compositions are always used. The diagrams can thus be understood as follows:

- Composition relationships: depicted as an arrow in these diagrams, with the source being a component (that is, and aggregate property) of the target.
- Sub-class relationships: depicted as an arrow with an empty diamond head; the source is a sub-class of the target, which is typically an abstract class (name of object is italicized)
- Non-composition, non-sub-class relationships are shown with simple lines.

The high-level model is shown below, and each major section is described in more detail in the following sections.

Note that all text properties holding natural language strings should be repeatable with indication of the language and the string of which others are translations. Similarly, controlled vocabularies and other coded representations are assumed to include information about the identification, version, and ownership of the source of the coded values of controlled terms, but this detail is not reflected in the conceptual model, being judged an aspect of implementation.

Notes are assumed to have information about their type, and also the language in which they are written. These details are not included in the conceptual model at this stage.

---

[1] Statistical Data and Metadata eXchange. http://sdmx.org/

[2] Resource Description Framework. http://www.w3.org/RDF/

[3] Unified Modeling Language

Related materials are also likely to be implemented as structured properties, including full citation information and typing, but they are depicted here as simple properties at the conceptual level.

Note that while version numbers are provided as properties of many objects, these are descriptive – the model is not intended to support management of the metadata objects themselves, but only searching and discovery of those objects. Thus, the version properties of objects reflect the harvested/reported version as stored in the portal, but the version relationships do not reflect any particular versioning strategy. They do not reflect the version of the registration at the portal, but only the reported version of the object.

The metadata model presented here does not include those fields, which will be required for implementation and management of the registrations themselves.

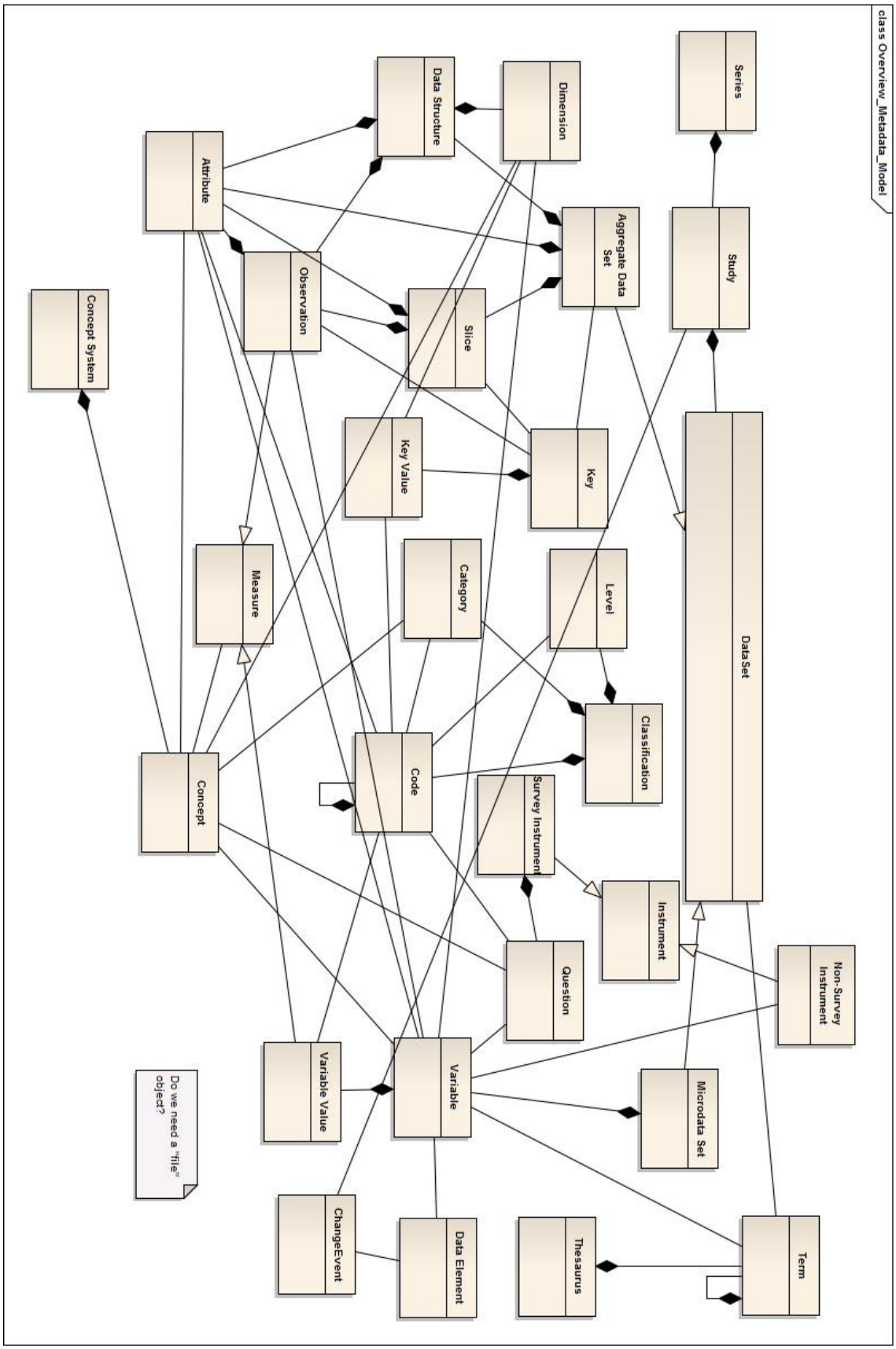


**Figure 1: Complete high-level object model**

## III. Series, Studies, and Data Sets

Each *Series* could have a set of *Studies*, and represents longitudinal and repeat cross-sectional data collection, and could also represent the on-going data collection practiced by NSIs[4], where each wave of collection is a *Study*. Each study can have one or more data sets associated with it [Figure 2].

Note that there is an abstract Data Set, which is the super-class for the aggregate data set and the microdata set. This modelling would support searches for data sets of any type, whether aggregate or microdata.

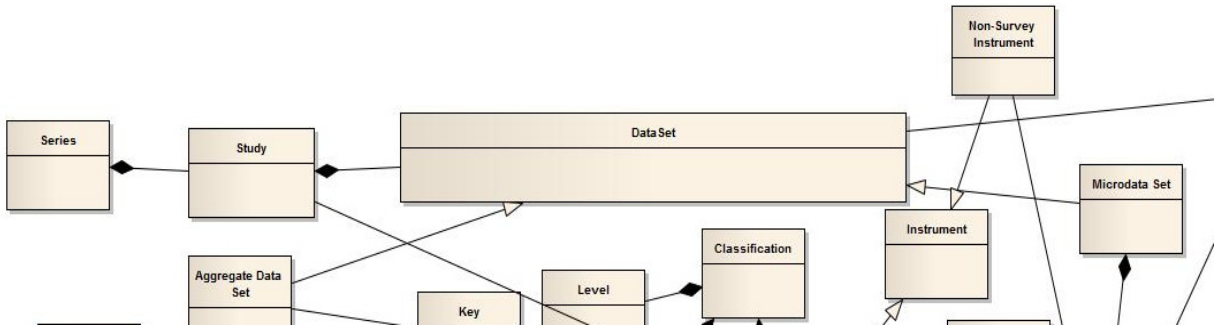


**Figure 2: Series, Studies & Data Sets.**

### A. Object Properties: Series

A series represents repeat data collection, across time, geography, population, or other dimension. It is in effect a group of related studies.

**Id** – A unique identifier for the object. *Required*.
**Agency** – The owner of the object. *Required*.
**Version** – The version of the object. Defaults to "version 1.0" if not specified.
**Title** – A natural-language title for the series. May be repeated for title translations. *Optional.*
**Subtitle** - A natural-language sub-title for the series. May be repeated for translations. *Optional.*
**Alternative title** – An alternative natural language title for the series, to support the WP5 metadata model. May be repeated for translations. *Optional.*
**Abstract** – A natural-language abstract for the series. May be repeated for translations. *Optional.*
**Topical keywords** – Terms to assist in searching for the object, typically coming from a controlled vocabulary. *Optional and repeatable.*
**Geographic coverage, coded** – A description of geographic coverage as indicated with a code coming from a geographic classification. Describes the super-set geography of the series across all studies/waves. *Optional and repeatable.*
**Geographic coverage, textual** – A textual description of the geographic coverage of the series. Describes the super-set of geographical coverage across all studies/waves. Repeatable for translation. *Optional.*

[4] National Statistical Institutes

**Reference period** - This is the time period of the measured phenomenon (temporal coverage), describing the earliest and latest dates covered by any of the studies in the series. It is represented as an ISO date/time[5], which supports the description of periods. *Optional.*
**Date of collection** - This is the point in time when the earliest and latest data in any study in the series was collected. Represented as an ISO date/time, which supports the description of periods. *Optional*.
**Frequency of data collection** – This is the description of the frequency of data collection for the studies within the series, expressed in a natural language. May be repeated for translations. *Optional.*
**Time method** – This is a description of the time methodology used within the studies associated with the series, expressed in a natural language. Repeatable to support translations. *Optional.*
**Data producer** – This is the natural-language name of the data creator (not necessarily the name of the archive or statistical agency which disseminates the data). Repeatable to support translations. *Optional.*
**Change event –** This documents changes in methodology between two instances of the same study within a series. *Optional and repeatable.*
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable.*

### B. Object Properties: Study

A study represents a single wave of data collection. This can be data collection across geographies or not, but is bounded by being a data collection for a specific period or point in time.

**Id** – A unique identifier for the object. *Required.*
**Agency** – The owner of the object. *Required*.
**Version** – The version of the object. Defaults to "version 1.0" if not specified.
**Title** – A natural-language title for the study. May be repeated for title translations. *Optional*.
**Subtitle** - A natural-language sub-title for the study. May be repeated for translations. *Optional*.
**Alternative title** – An alternative natural language title for the study, to support the WP5 metadata model. May be repeated for translations. *Optional*.
**Series reference** – A reference to a series of which the study is a part. *Optional and repeatable*.
**Abstract** – A natural language abstract for the study. May be repeated for translations. *Optional*.
**Topical keywords** – Terms to assist in searching for the object, typically coming from a controlled vocabulary. *Optional and repeatable*.
**Geographic coverage, coded** – A description of geographic coverage as indicated with a code coming from a geographic classification. Describes the super-set geography of all variables or other data within the study. *Optional and repeatable*.
**Geographic coverage, textual** – A textual description of the geographic coverage of the series. Describes the super-set of geographical coverage across all variables or other data within the study. Repeatable for translation. *Optional*.
**Reference period** - This is the time period of the measured phenomenon (temporal coverage), describing the earliest and latest dates covered by any of the variables or observations in the study. It is represented as an ISO date/time, which supports the description of periods. *Optional*.

[5] http://www.iso.org/iso/catalogue_detail?csnumber=40874

**Date of collection** - This is the point in time when the earliest and latest data in the study was collected. Represented as an ISO date/time, which supports the description of periods. *Optional*.
**Frequency of data collection** – This is the description of the frequency of data collection for the study, expressed in a natural language. May be repeated for translations. *Optional*.
**Time method** – This is a description of the time methodology used within the study, expressed in a natural language. Repeatable to support translations. *Optional*.
**Data producer** – This is the natural language name of the data producer (not necessarily the name of the archive or statistical agency which disseminates the data). Repeatable to support translations. *Optional*.
**Universe** – A description of the universe of the study, in a natural language. Repeatable to support translation. *Optional*.
**Population size** – The size of the population covered in the study, expressed as an integer. Included to support the WP5 model. *Optional*.
**Sample size** – The size of the studies' population, expressed as a fraction of the population size property. *Optional*.
**Kind of Data** – The kind of data collected by the study, described in a natural language. Repeatable for translations. *Optional*.
**Methodology Description** – A natural language description of the methodology for the study, including aspects of data collection and study design. Repeatable to support translation. *Optional*.
**Funding information** – Information relevant to the funding of the study, expressed in natural language. Repeatable to support translation. *Optional*.
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable*.
**Related materials** – A reference to a document or other source of information that adds additional information about the study. *Optional and repeatable*.
**Data source** – A link to a questionnaire or other description of the data source. Included to support the WP5 model. *Optional*.
**Source publications** – A link to a publication relevant to the data source. Included to support the WP5 model. *Optional*.

### C. Object Properties: Data Set (Abstract)

The data set object represents a file of data. It is an abstract class that is sub-classed into aggregate and micro-data/unit record data set objects, which inherit all its properties. Access to data is controlled at the data set level – any given study may have data sets with varying levels of access for different types of users, etc.

**Id** – A unique identifier for the object. *Required*.
**Agency** – The owner of the object. *Required*.
**Version** – The version of the object. Defaults to "version 1.0" if not specified.
**Title** – A natural language title for the data set. May be repeated for title translations. Maps to Name property in WP5 model. *Optional*.
**Type of User** – This field provides a term describing the type of user, intended to help control access to the data set itself. This is a textual field, but is not subject to translation. Defaults to "Public". Included to support the WP5 model. *Optional.*

**Type of Access** – This field provides a term describing the type of access, intended to help control access to the data set itself. This is a textual field, but is not subject to translation. Defaults to "Open". Included to support the WP5 model. *Optional.*
**Cost of Access** – This field provides a term describing the cost of access to the data ("Free", etc.) intended to help control access to the data set itself. This is a textual field, but is not subject to translation. Defaults to "Free". Included to support the WP5 model. *Optional*.
**Access Conditions** - This is a natural-language description of the conditions of access. Repeatable for translation. Included to support the WP5 model. *Optional*.
**Contact** – Contains natural-language contact information regarding the data file. Repeatable for translation. Included to support the WP5 model. *Optional*.
**Description** - A description of the data file in natural language. Repeatable to support translation. *Optional*.
**Type of data** – A term describing the type of data (e.g., register sample, register census, survey sample, survey census, mixed). *Optional and repeatable.*
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable.*

[Note, we do not wish to include the following fields but we maybe required to do so:
Topical keywords
Geographic coverage coded
Geographic coverage textual
Reference period (time of measured phenomenon or temporal coverage).
Date of collection
Frequency of data collection
Time method
Data producer]

### D. Object Properties: Aggregate Data Set

This object inherits from the Data Set object, and contains multi-dimensional data typically expressed as an SDMX data set or a DDI[6] NCube[7]. (See section IV, below.)

**Slice –** A subset of the aggregate data set which wild-cards one or more dimensions, describing a regular section of the possible Cartesian "cube" of data. *Optional and repeatable*.
**Attribute –** A named, typed property that is the use of a concept. *Optional and repeatable*.
**Data Structure –** Each aggregate data set has a Data Structure, which describes its dimensionality, representations, attributes, and uses of concepts (see below).
**Key** – A key within the data set that is associated with one of more actual observation values.

### E. Object Properties: Microdata Set

This object inherits from the Data Set object, and contains unit-record data typically described in DDI, but potentially described in SDMX.

**Variable –** Variables hold the values for each case/unit record in the microdata set, as a typed value. *Required and repeatable*.

---

[6] Data Documentation Initiative http://www.ddialliance.org/

[7] "Describe the logical structure of an n-dimensional array, in which each coordinate intersects with every other dimension at a single point. The NCube has been designed for use in the markup of aggregate data." http://www.ddialliance.org/bp/definitions

## IV. Aggregate Data Sets

This portion of the model shows a simplified version of the SDMX model: each aggregate data set would have dimensions, attributes, and measures, each associated with a concept, and described with a data structure object [Figure 3]. A data set itself will have the values for each dimension as key values, which compose keys for each slice or observation. Dimensions are ordered. A Slice is either a time series or a cross-section – this represents the indicator, which may often be the target of a search. It is possible for key values or attributes to have coded representations. Relationships to concepts should be considered here, as these could be searched through the data structures, rather than as direct properties of the instance objects themselves. Note key values might not relate directly to concepts.

Note that this structure could describe either an SDMX data set or a DDI NCube.

Source links to microdata may not be needed for data discovery, but should be supported in the model. How this link could be populated may need to be discussed in WP12 reports.

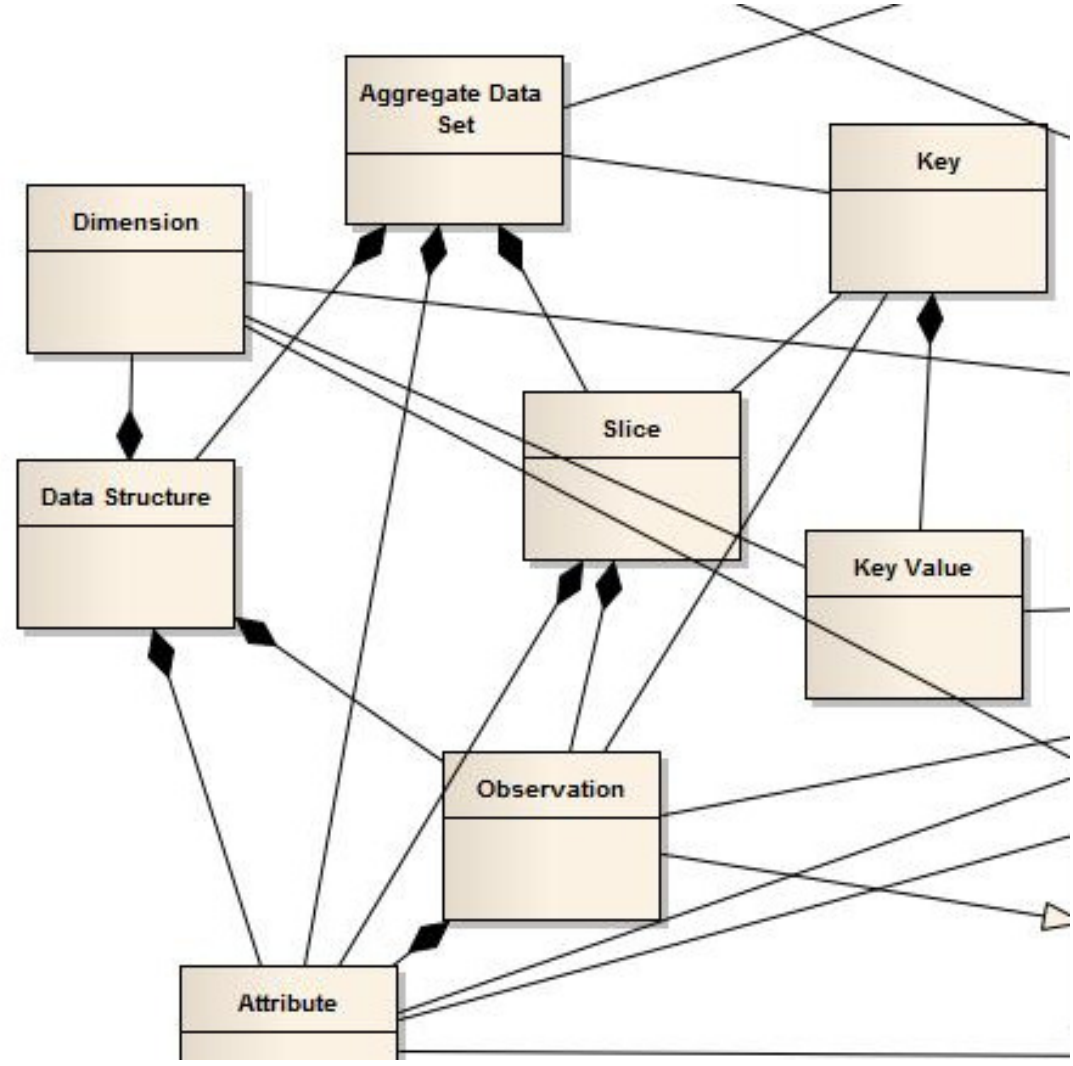


**Figure 3: Aggregate Data Sets.**

### A. Object Properties: Dimension

This is the structural description of a dimension used to define a multi-dimensional cube of data. This corresponds to an SDMX or DDI dimension.

**Id** – A unique identifier for the object. *Required*.
**Name** - A natural language identifier for the dimension. May be the same name as the concept (as in the case of all SDMX dimensions). Repeatable to support title translations. *Optional*.
**Description –** A natural language description of the dimension. Repeatable to support translations. *Optional*.
**Concept** – The statistical concept used as the dimension. *Optional*.
**Variable** – The variable in a microdata set used to populate the dimension in an aggregate data set in that particular tabulation. *Optional*.
**Representation** – The value type of the dimension, typically a coded value taken from a codelist, but may also be an identifier, a unique number, or a unique string. *Required*.
**Role** – An indication of the role of the dimension, if geographic, time, frequency (periodicity), or other specialized role as defined in the SDMX or DDI models.
**Rank –** This is an integer specifying the ordering of the dimension within the key, starting at 1 and incrementing.

### B. Object Properties: Data Structure

This object holds the description of the structure of a cube of data, and corresponds to an SDMX data structure definition or a DDI NCube structure.

**Id** – A unique identifier for the object. *Required*.
**Agency** – The owner of the object. *Required*.
**Version** – The version of the object. Defaults to "version 1.0" if not specified.
**Name** - A natural-language identifier for the data structure. Repeatable to support title translations. *Optional*.
**Description –** A natural-language description of the data structure. May be repeatable for translations. *Optional*.

### C. Object Properties: Attribute

This object represents a descriptive, non-identifying concept associated with an aggregate data set, a slice, or an observation. It corresponds to the attribute object found in the SDMX and DDI NCube models.

**Id** – A unique identifier for the object. *Required*.
**Name** - A natural language identifier for the attribute. May be repeated for title translations. May be the same name as the concept (as in the case of all SDMX attributes). *Optional*.
**Description –** A natural language description of the attribute. Repeatable to support translations. *Optional*.
**Concept** – The statistical concept used as the attribute. *Optional*.
**Variable** – The variable in a microdata set used to populate the attribute in an aggregate data set in that particular tabulation. *Optional*.
**Representation** – The value type of the attribute - a coded value taken from a codelist, an identifier, a number, or a string. *Required*.

### D. Object Properties: Slice

A slice is a regular section of the full possible Cartesian product of a multi-dimensional data cube. This is represented by wildcarding one or more of the possible dimensions of the aggregate data set.

**Key** – A set of valid values for a valid set of dimensions, as specified in the related data structure. At least one of the allowed dimensions must have a wildcard indicator ("*") as a value, even though this may not be a valid value according to the description of the dimension. *Required and repeatable*.
**Attribute –** An attribute associated with the slice. *Optional and repeatable*.

### E. Object Properties: Observation

This is the statistical concept that represents the values of each observation. It does not hold the value of the observation, but merely the structural information about the observation. It inherits from the measure object, which provides information about the types of its value.

**Concept** – A statistical concept used to represent the observation value. *Optional*.
**Variable** – A variable in a microdata set used to populate the observation values of the aggregate data set in a tabulation. *Optional*.
**Attribute** – An attribute associated with the observation in a data structure. *Optional*.
**Key** – The identifying key of the observation, including a complete set of specified key values as permitted by the data structure. *Optional*.

### F. Object Properties: Key

This represents the keys found within a data set that are associated with actual data. This is to support the discovery of data within a "sparse" data cube, to identify which types of observations are available for a given key within the data set, and to describe slices.

**Key Value -** A valid key value as allowed by the data structure. *Required and repeatable*.

### G. Object Properties: Key Value

This object is a component of a key.

**Value** – A code or literal value, valid as described for the associated dimension in a data structure. *Required and repeatable*.
**Dimension** – The dimension associated with the key value, as allowed by the data structure. *Required*.

## V. Measures and Variable Values

This group of objects describes the actual values found in data sets, whether aggregate or microdata. This structure should support querying on each distinct type of object, either specifically or more generically [Figure 4].

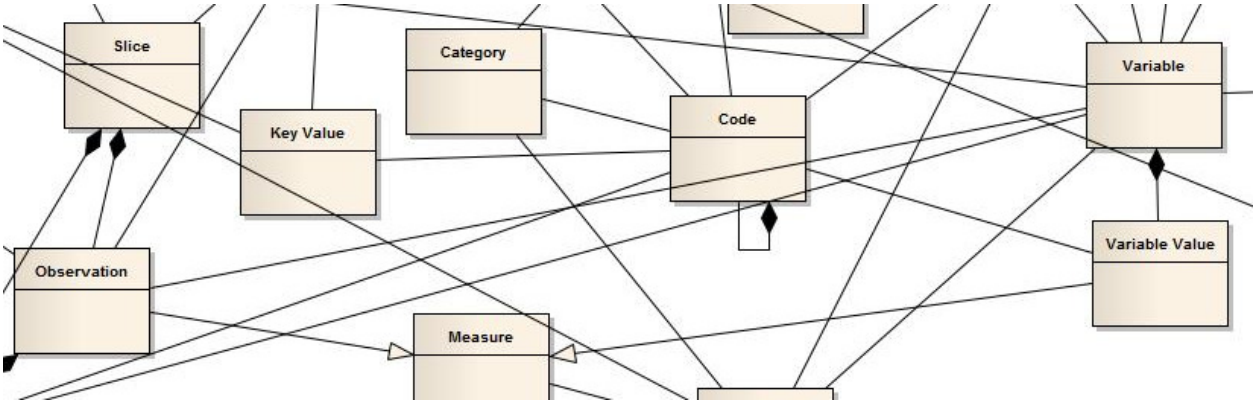


**Figure 4: Measures and Variable Values**

### A. Object Properties: Measure

This object represents the statistical measure (ratio, per cent, total amount, etc.) and is used to describe the types of observation values and variable values found in data sets. It is given meaning by the concept with which it is associated.

**Concept** – The concept that defines the measure.

### B. Object Properties: Variable Value

This is the structural description of the value of a variable. It inherits from a measure, which describes the concept identifying the measure of the variable value. This concept is distinct from the concept that defines the variable itself.

**Representation -** The value type of the dimension, typically a coded value taken from a codelist, but may also be an identifier, a unique number, or a unique string. *Required*.
**Level –** The level within a classification from which the variable takes its value, if the representation property indicates the use of a classification.

## VI. Concepts, Classifications, Codes, Categories, and Thesauri

These are a set of heavily reused objects. Concepts represent statistical concepts, which are different from thesauri [Figure 5] (the latter being *ad-hoc*, while the concepts themselves are more formally structured into concept systems.) There is an open question as to whether concepts and concept systems have a relationship to classifications – We would argue that they do, but this is not reflected in either SDMX or DDI today.

Classifications are very simplified here – they consist of a set of categories and codes, which themselves can reflect the structures of how they are used either in DDI or SDMX today. Note that levels are not represented. Codes are used by variables, questions, dimensions, attributes, *etc*. [Figure 5].

Thesauri represent collections of keywords, search terms, DDI controlled vocabularies, SDMX Category Schemes, *etc*. We have placed here the simplest representation of thesauri for discussion purposes [Figure 6].

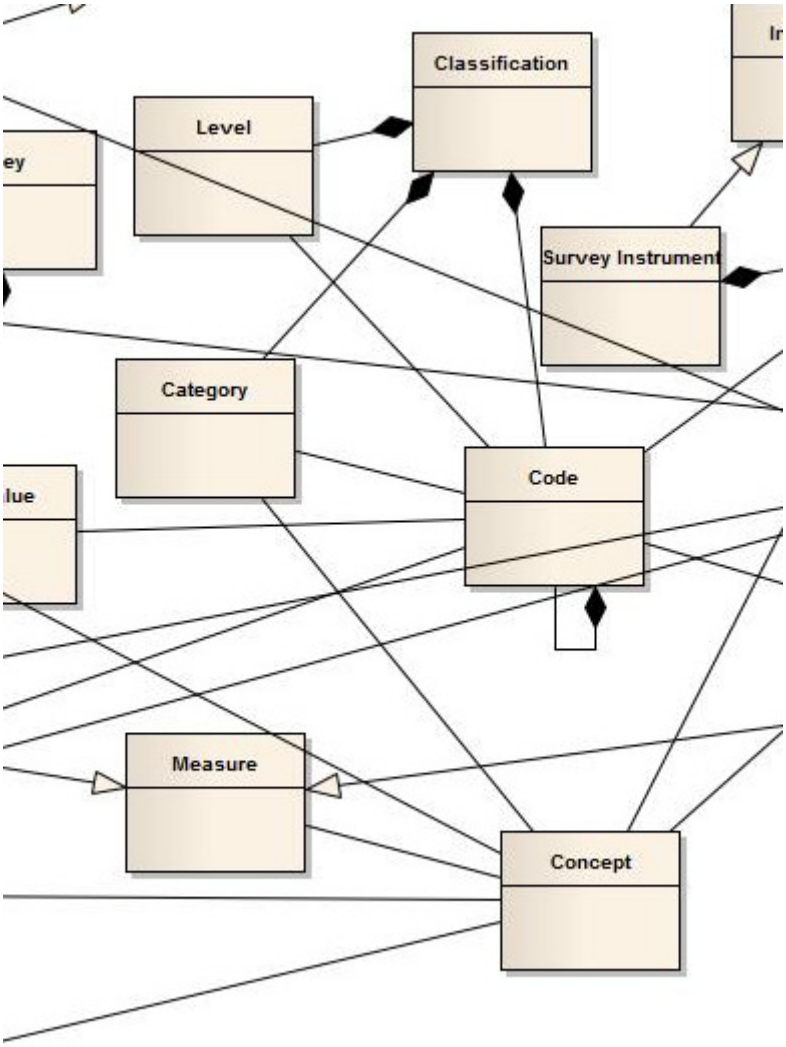


**Figure 5: Concepts, Classifications, Codes, and Categories.**

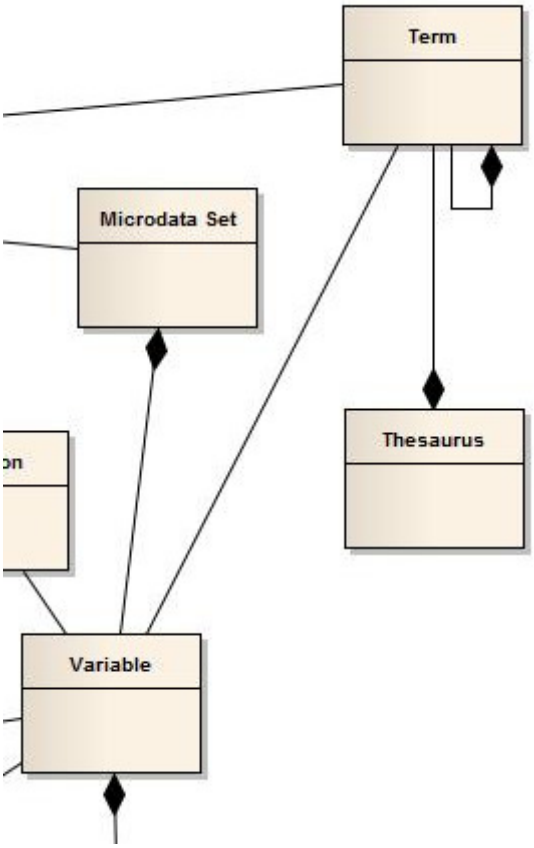


**Figure 6: Thesauri**

### A. Object Properties: Concept

This object represents a statistical concept, providing the formal name and definition, as opposed to associated search terms or keywords coming from a thesaurus.

**Id** – A unique identifier for the concept. *Required*.
**Agency** – The owner of the object. *Required*.
**Version** - The version of the object. Defaults to "version 1.0" if not specified.
**Name** – The formal natural-language name of the concept. Repeatable to support language translation. *Required*.
**Labels** – Other labels associated with this concept, typically coming from sources such as statistical processing systems that truncate or otherwise assign an alias to the concept. *Optional and repeatable.*
**Description** – The formal definition of the concept in a natural language. Repeatable to support language translation. *Required*.
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable*.
**Related materials** – A reference to a document or other source of information that adds additional information about the concept. *Optional and repeatable*.

### B. Object Properties: Concept System

**Id** – A unique identifier for the concept system. *Required*.
**Agency** – The owner of the object. *Required*.
**Version** - The version of the object. Defaults to "version 1.0" if not specified.

**Name** – The formal natural-language name of the concept. Repeatable to support translation. *Required*.
**Description** – A description of the concept system in a natural language. Repeatable to support translation. *Required*.
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable*.
**Related materials** – A reference to a document or other source of information that adds additional information about the concept system. *Optional and repeatable*.

### C. Object Properties: Code

**Id** – A unique identifier for the code. *Required*.
**Text –** The textual or numeric code as found within the data. *Required*.
**Parent Code** – A code that is the parent in a hierarchical coding scheme. *Optional*.
**Category** – A category that indicates the meaning of the code. *Required*.
**Level –** The level with which the code is associated within a levelled classification.

### D. Object Properties: Classification

**Id** – A unique identifier for the classification. *Required*.
**Agency** – The owner of the object. *Required*.
**Version** - The version of the object. Defaults to "version 1.0" if not specified.
**Name** – The formal natural-language name of the classification. Repeatable to support translation. *Required*.
**Description** – A description of the classification in a natural language. Repeatable to support translation. *Required*.
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable*.
**Related materials** – A reference to a document or other source of information that adds additional information about the classification. *Optional and repeatable*.

### E. Object Properties: Category

A category is a unit of meaning associated with a code.

**Id** – A unique identifier for the category. *Required*.
**Description** – The formal definition of the classification in a natural language. Repeatable to support translation. *Required*.

### F. Level

This object represents a single level within a levelled classification.

**Concept** – A concept that provides the meaning for the level. *Optional*.

### F. Object Properties: Thesaurus

A thesaurus is a set of terms used to help in searching, but is not a formal concept system. Examples include ELSST[8] and similar schemes.

**Id** – A unique identifier for the thesaurus. *Required*.
**Agency** – The owner of the object. *Required*.

[8] Multilingual European Language Social Science Thesaurus. http://elsst.esds.ac.uk/login.aspx

**Version** - The version of the object. Defaults to "version 1.0" if not specified.
**Name** – The formal natural-language name of the thesaurus. Repeatable to support translation. *Required*.
**Description** – A description of the thesaurus in a natural language. Repeatable to support translation. *Required*.
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable*.
**Related materials** – A reference to a document or other source of information that adds additional information about the thesaurus. *Optional and repeatable*.

### G. Object Properties: Term

A term is a reserved word used in a thesaurus.

**Id** – A unique identifier for the term. *Required*.
**Text** – The reserved word, in a natural language. Repeatable to support translation. *Required*.
**Definition** – The meaning of the term, expressed in a natural language. Repeatable to support translation. *Optional*.
**Parent term** – A parent term for use with hierarchical thesauri. *Optional*.

## VII. Variables, Data Elements, Questions, Change Events, Survey Instruments, and Instruments

This portion of the model represents some of the core objects of the model in terms of searches.

Variables are the key contents of microdata sets – they have a number of properties. The value of variables is modelled in some detail – the value itself can be either a classificatory value or a statistical value, as these have different functions when mapped against aggregate data. A classificatory value would be similar to a key value (that relationship should be added to the model) while the statistical value would be mapped to an observation in an aggregate data set. Relationships between variable values and how they function as independent or dependent variables in creating aggregate observations is not reflected here, although this may be important [Figure 7].

Note that logical record structures are not reflected here – these are seen as not useful for the purposes of search and discovery. However, comparability of variables across cycles of a data collection is important – to identify cases where variables with different identifiers are meant to be comparable is reflected through their relationship with the same data element.

Variables have relationships with questions, which are themselves grouped into survey instruments. Survey instruments are a specialization of a generic instrument object, as we may wish to capture information regarding non-survey data sources (e.g. administrative registers).
These structures may seem too detailed, but if we consider the different searches that might be made, this level of detail is seen as possibly necessary. For example, a researcher may only be searching on key values and classificatory variables, rather than the more typical search across all variables [Figure 7].

Not represented here are summary statistics, even though these will be very useful for display. We would model these as objects with relationships to variables, codes, and microdata sets.

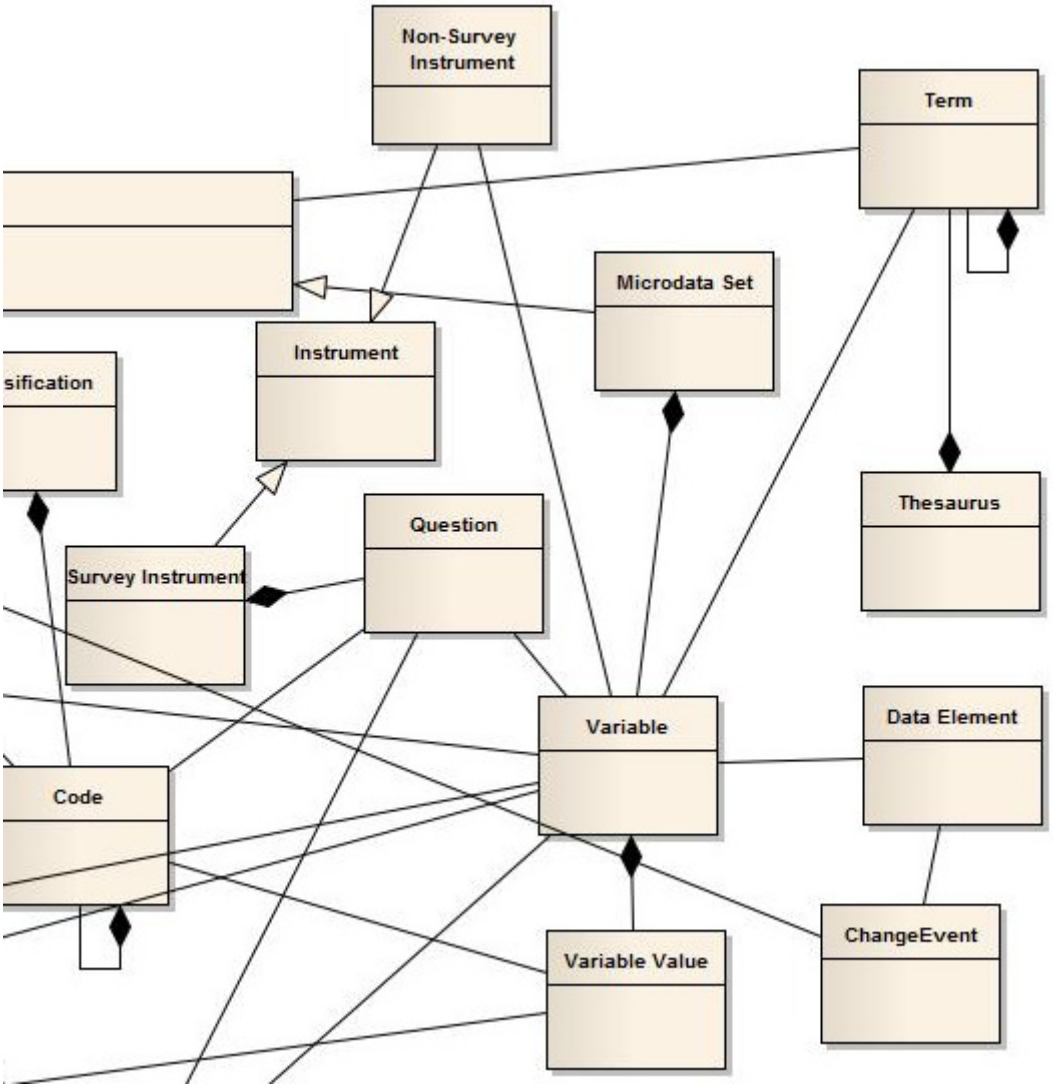


**Figure 7: Variables, Data Elements, Questions, Change Events, Survey Instruments & Instruments.**

### A. Object Properties: Variable

This object represents a single data item in a microdata set. It may be related to a data element if it represents the repeat collection of a comparable data item in a repeated cycle of data collection.

**Id** – A unique identifier for the object. *Required*.
**Version** - The version of the object. Defaults to "version 1.0" if not specified.
**Agency** – The owner of the metadata object. *Required*.
**Name** – The native (non-portal) id of the variable, non-unique and expressed in a natural language. Repeatable to support translation. *Optional*.
**Label –** A natural language name of the variable as used in systems such as statistical packages that truncate variables names or otherwise modify them. *Optional*.
**Description** – A natural language description of the variable. Repeatable to support translation. *Optional*.

**Question** – The question object that was used to populate the variable. *Optional*.
**Universe –** A textual description of the set of respondents who are measured by the variable. A sub-set of the universe of the study. Repeatable to support translation. *Optional*.
**Concept** – The concept that formally describes the statistical concept measured by the variable. *Optional*.
**Variable value** – This is the structural description of the value of the variable. *Required*.
**IsWeight** – A Boolean value indicating that the variable holds a weight. Defaults to "false" if not specified. *Required*.
**IsTemporal** – A Boolean value indicating that the variable holds a time value. Defaults to "false". *Required*.
**IsGeographic** – A Boolean indicating that the variable holds a geographic value. Defaults to "false". *Required*.
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable*.
**Related materials** – A reference to a document or other source of information that adds additional information about the variable. *Optional and repeatable*.

### B. Object Properties: Data Element

This object is used to relate comparable variables populated by different waves of data collection within a series.

**Id** – A unique identifier for the object. *Required*.
**Concept –** The concept that describes the statistical concept measured by the variables associated with the data element. *Optional*.
**Change event** – A change event describing the evolution of variables related to this data element.
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable*.
**Related materials** – A reference to a document or other source of information that adds additional information about the data element. *Optional and repeatable*.

### C. Change Event

**Id** – A unique identifier for the object. *Required*.
**Type** – A reserved term indicating the type of the change event. *Optional*.
**Text** – A natural-language description of the change event. Repeatable to support translation. *Required*.
**Source** – The object (variable or study) that has been changed by the event. *Optional and repeatable*.
**Target** – The object (variable or study) that has resulted from the change event. *Optional and repeatable*.
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable*.
**Related materials** – A reference to a document or other source of information that adds additional information about the data element. *Optional and repeatable*.

### D. Object Properties: Question

This object represents a question from a survey instrument.

**Id** – A unique identifier for the question. *Required*.

**Standard question set name** – A natural language name for a standard question set of battery from which the question is taken. Repeatable to support translation. *Optional*.
**Survey instrument** – The survey instrument in which the question appears. *Optional and repeatable*.
**Pre-Text** – The question pre-text in a natural language. Repeatable to support translation. *Optional*.
**Question text** – The text of the question in a natural language. Repeatable to support translation. *Required*.
**Post-Text** – The question post-text in a natural language. Repeatable to support translation. *Optional*.
**Representation** - The response domain of the question, often a coded value taken from a classification, but may also be an identifier, a unique number, or a unique string. *Required*.
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable*.
**Related materials** – A reference to a document or other source of information that adds additional information about the variable. *Optional and repeatable*.

### E. Object Properties: Instrument

This is an abstract object representing any instrument used to collect data. It is sub-classed by survey instruments and non-survey instruments.

**Id** – A unique identifier for the concept system. *Required*.
**Agency** – The owner of the object. *Required*.
**Version** - The version of the object. Defaults to "version 1.0" if not specified.
**Name** – The formal natural language name of the instrument. Repeatable to support translation. *Required*.
**Description** – A description of the instrument in a natural language. Repeatable to support translation. *Required*.
**Study** – The study that used the instrument to collect data. *Optional and repeatable*.
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable*.
**Related materials** – A reference to a document or other source of information that adds additional information about the instrument. *Optional and repeatable*.

### F. Object Properties: Survey Instrument

Type: Maintainable

**Id** – A unique identifier for the concept system. *Required*.
**Agency** – The owner of the object. *Required*.
**Version** - The version of the object. Defaults to "version 1.0" if not specified.
**Name** – The natural language ID of the survey. May be non-unique. Repeatable to support translation. *Optional*.
**Title** – The natural language title of the survey. Repeatable to support translation. *Optional*.
**Description** – A description of the concept system in a natural language. Repeatable to support translation. *Required*.
**Survey link** – A link to a PDF or other format containing the survey. *Optional and repeatable*.
**Mode** – A value from a controlled vocabulary indicating the mode of the survey. *Optional and repeatable for multi-mode surveys*.
**Note** – A typed, repeatable note in a natural language. *Optional and repeatable*.

**Related materials** – A reference to a document or other source of information that adds additional information about the survey instrument. Note that a link to a file containing the survey (such as a PDF) is not considered a related material, but has a special type of link. *Optional and repeatable*.

### G. Object Properties: Non-Survey Instrument

This represents a non-survey data source such as a register or other data collection device. It inherits all properties from Instrument and has a link to Variable.

## VIII. SUMMARY

This high-level model will serve as a useful starting point in considering what metadata we can capture and use to support the user stories, which will inform the development of more detailed metadata models for implementation. We have focused here only on the statistical metadata, corresponding to the OS Model that is the deliverable. Implementation will have this model as a basis, with changes to facilitate implementation.