Metadata Schema for Data Dictionaries

SCOPE Metadata Group

April 9, 2014

Background:

The SCOPE (Statistical Community of Practice and Engagement) project was initiated in 2009 as an inter-agency effort to consolidate resources and practices across the major US federal statistical agencies. A team focused on metadata was initiated in 2013. The effort described here is part of the team’s deliverables.

Parallel with SCOPE, the White House began an effort to make federal government data available through a single resource on the Web known as Data.Gov. As part of the effort to build a unified Data.gov, the federal agencies have been asked to standardize the format of certain metadata about the datasets they offer online, to enable public users and computer programs to find and use the data.

One element in the set of descriptors for a data set is a pointer to a data dictionary, if applicable. This specification is an effort to define the elements needed to describe the contents of a data dictionary.

Membership:

Dan Gillman (BLS) – chair

Paul Bugg (OMB)

San Cannon (FRB)

Joe Dalaker (Census)

Mark Elbert (EIA)

Tim Kearley (BJS)

Peter Meyer (BLS)

Mary Moulton (BTS)

Kimberly Noonan (NSF)

Bill Savino (Census)

Marilyn Seastrom (NCES)

Joy Sharp (BTS)

Ann Summa (NASS)

Bruce Taylor (NCES)

Michael Valivullah (NASS)

Usage:

In this specification, dimensions for cross-tabulated data are treated as variables. Many other standards and specifications for describing data, such as DDI and SDMX, treat dimensions as separate objects. This decision was made to keep the number of elements to a minimum.

Notes:

1. An entry in a data dictionary is a description of a variable.
2. A Value Domain is a set of allowed values for a variable.
3. Multiplicity tells the reader of this document how many instances of a particular element are allowed in each entry.
4. All the fields specified below are assumed to be text for human readability. Interoperability and use by automated tools requires new considerations beyond the scope of this document.
5. Examples are included at the end of this document.

Definitions:

The definitions below are given to ensure unambiguous interpretation of this specification.

1. Datatype – The computational model for some data, characterized by axioms and operations, and containing a set of distinct values
2. Definition – Representation of a concept by a textual description used to convey its meaning
3. Description – Statement (possibly formal), account, or picture that explains or delineates
4. Designation – Representation of a concept by a signifier which denotes it
5. Dimension – Characteristic used to partition the elements of a universe into mutually exclusive and exhaustive subsets (e.g., sex partitions people into male and female)
6. Identifier – Kind of Name, for which its intended use is dereferencing the designated object
7. Name – Representation of an object by a signifier which denotes it
8. Permissible Value – Element of a Value Domain, and in the enumerated case is represented as an ordered pair consisting of a designation and meaning
9. Sub-Dimension – Dimension for another Dimension (e.g., counties sub-divide each state in the US)
10. Universe– The total membership of a defined class of, for example, people, objects, or events
11. Units of measure – Magnitude for some quantity, either physical or immaterial
12. Value domain – Set of allowed values for a variable
13. Value domain: enumerated – Value domain for which the allowed values are given as a list
14. Value domain: described – Value domain for which the allowed values are delineated by description
15. Variable – Map from the elements of a Universe to the elements of a Value Domain, where for each element of the Universe at most one element of the Value Domain is assigned

**Required Fields**

| **Field** | **Definition** | **Usage Comments** |
| --- | --- | --- |
| Identifier | Globally unique identifier for each data dictionary entry  Multiplicity [1..1] | Given the current emphasis on linked open data, this identifier should be a URI.  A data dictionary entry is a description of a variable. |
| Name of Variable | Linguistic expression designating the entry  Multiplicity [1..N] | More than one name for a data dictionary entry is permitted. |
| Definition of Variable | Representation of the meaning of a variable by a textual description used to convey its meaning Multiplicity [1..1] | Multiple definitions are permitted if they are in different languages. The assumption is one definition per language. |
| Type of Variable | The kind of entry provided  Multiplicity [1..1] | The allowed values are: Basic, Cross-tabulated, Derived, and Dimensional. |
| Name of Value Domain | Linguistic expression designating the value domain  Multiplicity [1..N] | More than one name is permitted for a value domain. |
| Type of Value Domain | The kind of value domain  Multiplicity [1..1] | Value domains come in 2 types:  Enumerated – allowed values are specified as a list using the Permissible Value field below  Described – allowed values are specified in a rule or expression using the Description field below |
| Datatype | Allowed computations (i.e., the computational model) for data under this variable  Multiplicity [1..1] | In statistics, the allowed families of datatypes are Text, Nominal, Ordinal, Interval, and Ratio. Additional datatypes may be required to specify other data. |

**Required-if-Applicable Fields**

| **Field** | **Definition** | **Usage Comments** |
| --- | --- | --- |
| Universe | The total membership of a defined class of people, objects, or events Multiplicity [0..1] | This is not applicable to Dimensional variables. It is required otherwise. All adult persons in the US is an example of a universe. |
| Permissible Value | Pair consisting of a designation and meaning; and one of the allowed values for a variable  Multiplicity [0..N] | Used for Enumerated type Value Domains – The pair <M, male> might be a way to express the idea that the letter M is used to denote the male sex. The word “male” contains the meaning associated with the usage of the letter “M” in some database entry. In this sense, it is a permissible value. |
| Description | Rule or expression for specifying permissible values  Multiplicity [0..1] | Used for Described type Value Domains – The description might be the set of real numbers between 0 and 1 in Arabic numerals. This choice is taken when listing the allowed values is not feasible, helpful, or instructive to list them individually. |
| Unit of Measure | Units used to quantify values  Multiplicity [0..1] | Degrees Celsius is a unit of measure for temperature. Miles per hour is a unit of measure for speed. |
| Precision | Smallest unit in which numeric data are reported  Multiplicity [0..1] | Temperatures measured to the nearest one hundredth of a degree have a precision of hundredths. Temperatures measured to the nearest whole degree have a precision of units. |
| Dimensions | List of names of dimensions for cross-tabulated variables  Multiplicity [0..N] | This element is used only in the description of cross-tabulated variables. Used to name (i.e., reference) the description of a dimensional variable. See Type of Variable field. |
| Sub-Dimension | A dimension that is a further refinement of another dimension  Multiplicity [0..1] | Used in the description of a dimensional variable that is used as a sub-dimension. Value in this field is the name (a reference) to the related parent dimension.  Counties in the US are a sub-dimension of the states. Each level of a hierarchical classification (e.g., NAICS) is a sub-dimension of the level above. |
| Statistic | Particular statistic a derived or cross-tabulated variable represents  Multiplicity [0..1] | A mean (or average) is derived from other data. This is a statistic. This element only applies to derived or cross-tabulated variables. Even some derived variables do not represent statistics, such as classifications whose values are based on combinations of others. In statistics, these are called re-codes. |

**Examples of Describing Variables**

Here are 4 variables described as if they are in a microdata file:

* 1. Name – sex
  2. Definition – Sex of a person
  3. Universe – US people in 2014
  4. Type – basic
  5. Value domain
     1. Name – sex codes
     2. Type – enumerated
     3. Permissible values
        1. <m, Male sex>
        2. <f, Female sex>
        3. <o, Other – unspecified>
     4. Datatype – nominal
  6. Name – state
  7. Definition – state of residence of person
  8. Universe – US people in 2014
  9. Type – basic
  10. Value domain
      1. Name – postal state codes
      2. Type – enumerated
      3. Permissible values
         1. <AL, state of Alabama>
         2. <AK, state of Alaska>
         3. …
      4. Datatype – nominal
  11. Name – county
  12. Definition – county of residence of person
  13. Universe – US people in 2014
  14. Type – basic
  15. Value domain
      1. Name – FIPS county codes
      2. Type – enumerated
      3. Permissible values
         1. <001, Autauga county in Alabama>
         2. <003, Baldwin county in Alabama>
         3. …
      4. Datatype – nominal
  16. Name – income
  17. Definition – income from all sources for person
  18. Universe-US people in 2014
  19. Type – basic
  20. Value domain
      1. Name – monetary amount
      2. Type – described
      3. Description – non-negative number in Arabic numerals

Here is the description of a table based on these variables:

Dimensions

* 1. Name – sex names
  2. Definition – Sex of a person
  3. Type – dimensional
  4. Value domain
     1. Name – sex names
     2. Type – enumerated
     3. Permissible values
        1. <male, Male sex>
        2. <female, Female sex>
        3. <other, Other – unspecified>
  5. Name – state names
  6. Definition – state of residence of person
  7. Type – dimensional
  8. Value domain
     1. Name –state names
     2. Type – enumerated
     3. Permissible values
        1. <alabama, state of Alabama>
        2. <alaska, state of Alaska>
        3. …
  9. Name – county names
  10. Definition – county of residence of person
  11. Type – dimensional
      1. Sub-dimension to – state names
  12. Value domain
      1. Name – county names
      2. Type – enumerated
      3. Permissible values
         1. <Autauga - AL, Autauga county in Alabama>
         2. < Baldwin - AL, Baldwin county in Alabama>
         3. …

Cross-tabulated variable

* 1. Name – income totals
  2. Definition – income from all sources for person
  3. Universe – US people in 2014
  4. Type – cross-tabulated
     1. Statistic – total
     2. Dimensions
        1. sex names
        2. state names
        3. county names
  5. Value domain
     1. Name – monetary amount
     2. Type – described
     3. Description – non-negative number in Arabic numerals
     4. Precision – 2 decimals
     5. Unit of measure – US dollars
     6. Datatype – ratio