Metadata Element Set for On-Line Data Dictionaries

Purpose: To provide a set of elements for describing variables in on-line data dictionaries.

Background: Since the advent of the Internet and the Web, it has been the desire of the US federal statistical agencies to make all their data available to the public in a unified fashion. Currently, each agency maintains its own web site, and users are forced to know which agency produces what kinds of data. The aim of the unified approach is to provide a “one-stop shop” access to statistical data.

The Data.Gov initiative of the current administration has the same goal. However, the scope covers all federal agencies, not just the statistical ones. Now, in support of Data.Gov, the Open Data Project is leading the way to establish a catalog of data sets from every agency. A metadata scheme has been developed for this purpose. One of these elements is a placeholder for a data dictionary. The elements below in this document can be used for that purpose. In any case, in order to be able to read and process structured data in data sets, one must have a data dictionary along with the data.

The DCAT vocabulary (<http://www.w3.org/TR/vocab-dcat/>), a W3C recommendation, is widely used for catalogs of data sets. From the DCAT web site – “DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. This document defines the schema and provides examples for its use.” A catalog for which the underlying metadata model conforms to DCAT is interoperable with any other such catalog. However, DCAT does not address the data dictionary that must go along with any structured data set.

Currently, there is no agreed upon specification for the contents of data dictionaries. Leaving this problem for each agency to solve just creates the problem anew. As long as two data dictionaries are structured differently, some scheme for understanding each of them is necessary. This just recapitulates the problem: first the problem exists for the data themselves, then it exists for the data dictionary. Until there is some standard specification, each description will have to be described itself.

Solution: To avoid this problem, a widely agreed upon specification for describing data dictionaries is recommended. This document provides the elements for achieving that.

The elements provided describe a variable. The number of allowed occurrences of each element is provided, where the [x..y] notation means x is the minimum and y is the maximum allowable number of occurrences. The use of “n” as a maximum means the number of occurrences is unbounded. An element listed as required under an element listed as optional is conditional upon the optional element being populated.

The element set provided below is conforms to the subset of the model describing variables provided in the Generic Statistical Information Model (GSIM), built in 2012 by a set of experts from statistical offices around the world, including the US. Information about GSIM can be found at <http://www1.unece.org/stat/platform/pages/viewpage.action?pageId=59703371>.

The element set provided describes a variable, but it may be expanded to include descriptions of all the variables for some survey or agency by normalizing it. GSIM provides such a normalized model, and we recommend its use for that purpose.

Definitions of the elements and two examples follow the general listing of elements below.

General Elements:

* Variable
  + Name [1..n]
  + Definition [1..1]
  + Unit Type [1..1]
    - Name [1..1]
    - Definition [1..1]
  + Value Domain [1..1]
    - Name [1..1]
    - Enumerated [0..1] Either this or Described
      * Category [1..n]
        + Designation [1..1] Usually a code
        + Definition [1..1]
    - Described [0..1] Either this or Enumerated
      * Description [1..1] Description of values
      * Designation [1..1] Representation of values
  + Datatype [1..1]
  + Units of Measure [0..1]

Definitions:

Many of the definitions provided below are taken from GSIM.

1. Category – A concept whose role is to extensionally define and measure a variable
2. Datatype – The computational model for some data, characterized by axioms and operations, and containing a set of distinct values
3. Definition – Representation of a concept by a textual description used to convey its meaning
4. Description – Statement (possibly formal), account, or picture that explains or delineates
5. Designation – Representation of a concept by a sign which denotes it
6. Name – Label for some object
7. Unit type – The total membership of a defined class of people, objects, or events
8. Units of measure – Magnitude for some quantity, either physical or immaterial
9. Value domain – Set of allowed values for a variable
10. Value domain: enumerated – Value domain for which the allowed values are given as a list
11. Value domain: described – Value domain for which the allowed values are delineated by description

Example 1:

* Variable
  + Sex
  + Biological sex of a person (not the gender)
  + Unit Type
    - person
    - people alive at the time of observation
  + Value Domain
    - Sex codes
    - Enumerated
      * <F, Female>
      * The female sex
      * <M, Male>
      * the male sex
  + Datatype – Nominal
  + Units of measure – N/A

Example 2:

* Variable
  + Age
  + Age in whole years of a person since last birthday
  + Unit Type
    - person
    - people alive at the time of observation
  + Value Domain
    - Count of years
    - Described
      * Natural numbers (i.e., 0, 1, 2, 3, …)
      * Represented by Arabic numerals
  + Datatype – Interval
  + Units of Measure – whole years