Analysis: Describing a Data Cube in DDI4

# Introduction

This analysis of the capabilities of the DDI4 model is devoted to what is needed to describe a logical data cube. Currently, the model can describe rectangular data logically. We need to add the capability of describing cubes logically.

There are close ties between cube data and the underlying data (usually microdata) used to derive the cube. Our purpose is to describe a cube independently of the underlying microdata. Later work will make the links explicit.

# Dimensions and Cubes

A data cube is way to logically organize multi-dimensional data. It is not a presentation format, such as a table. In fact, if the number of dimensions in a cube is larger than 3, it is not possible to represent a cube visually. Here, we lay out the requirements for logically describing cubes. In so doing, we illustrate areas where DDI4 is missing constructs that are needed.

Dimensional data are those that include categories from several classification schemes (the dimensions) in order to specialize the universe for some underlying data. Then, some measure is applied to that specialized universe. An example might be average weekly wages of married females living in Ontario, Canada in March 2019. Here, the dimensions are Canadian provinces and gender. The cube measure is average weekly wage.

It is possible to have many dimensions in a cube, but for any one, each cell of the cube is defined by one category from each of the several dimensions. Sometimes, this combination is used as an identifier for each cell, but it is more than that. Consider the combinations produced by Canadian province, gender, and marital status (considered as single or married):

* Alberta, male, single
* Alberta, male, married
* Alberta, female, single
* Alberta, female, married
* British Columbia, male, single
* British Columbia, male, married
* British Columbia, female, single
* British Columbia, female, married
* ….
* Saskatchewan, male, single
* Saskatchewan, male, married
* Saskatchewan, female, single
* Saskatchewan, female, married

Each of these category combinations supplies a meaning, or semantic, to the cell to which it corresponds, and it shows how groupings are defined. This is the essential importance of the category combinations.

# Measures

Cube measures are (almost always) the quantitative variables that supply the numbers that appear in the cells of cubes. In our example above, the measure is average weekly wage. In microdata, a variable links each member of a population to some value (category or number). However, for dimensional data, the input to the variable is no longer the individual members of a population, rather they are the groups of those members defined by the category combinations (or cells). In this way, the variable associated with a cell has changed character. It is not a variable in the traditional sense, and in our example, the category combinations listed are the groupings.

In the example above, weekly wage is a variable in the underlying microdata. However, for each cell in the cube, the statistic “average” is applied to weekly wage for all members that meet the categorical criteria of the cell. The average is applied to the cell, i.e., the group, not the individuals assigned to the cell.

For other cube measures, other statistics are applied, e.g., totals, means (averages), quartiles, variance, or standard errors. Sometimes, for each cell a statistic produces more than one output. Sometimes, such as for covariance, the statistic needs more than one input.

# Cube Data Layout

We can use the Tall data structure to represent the content of cells in a cube. Content comprises the value(s) of the measure and the categories defining each cell. This is not illustrated here as other recent MRT position papers cover this topic.

# What Is Missing?

Here is a listing of features seen as missing from DDI4 that nevertheless are needed to adequately describe a cube. This listing does not consist of recommended changes to the model. Rather, these are the ideas that need to be conveyed. The list, in no particular order, follows:

1. The variable cascade does an excellent job of describing microdata, and the same structures can be used to describe variables used as measures in cubes. However, the distinctions needed to separate a domain of individuals (universe for microdata) from one of groups of individuals does not exist. There are terminological considerations that can be applied, but they are probably just necessary and not sufficient.
2. A logical description of a cube doesn’t exist. We need to link a cube to its dimensions and its measure(s). The dimensions need to be explicitly described so that we know which categories are included. For instance, NAICS has several levels of categories. Naming NAICS as a dimension is not sufficient. Depending on which level is picked, a different set of categories is defined (and a different number of cells in the cube results!). The ordering of the categories for each dimension is not important. The uniqueness of a cube does not depend on orderings of categories nor on that of the ordering of the dimensions themselves. Cubes depend on the measures and statistics used to produce them, an overall universe (say people or business establishments), the dimensions, and the particular categories in each dimension (possibly expressed as the levels in the dimensions).
3. A Viewpoint currently contains 3 roles: identifier, attribute, and measure. An additional role for cubes, dimension, needs to be added for those microdata categorical variables that are used to group units for cells. The value domains for the categorical variables identified as taking the dimension role are dimensions.
4. The size of a cube, i.e., the number of cells, is the product of the number of categories from each dimension. This may be obvious but is important metadata to convey about a cube.
5. Measures not only need a better description of their domains, we also need to know the statistic used – the derivation. We have the necessary structures in place, and may even have the requisite relationships in the model, but this has to be confirmed. Specifying the underlying (microdata) measure and statistic applied to produce a cube measure for a dimensional data cube is vital. It will not only convey meaning to the cube measure it will help determine whether a table produced from the cube is additive or not. Indices and rates, for instance, are not additive. Totals are additive. Percent is sometimes.
6. Additivity is a necessary consideration. Since index, rate, percent, total, etc. are similar to datatype families as described in ISO/IEC 11404 – General purpose datatypes, then it is through the datatype of the cube measure whether a cube is additive or not.
7. Tables often are conveyed as cubes embedded in other cubes. For example, take a cube defining population totals by age, race, sex, and marital status. Often, tables representing this cube also have totals with dimensions taken away, say only age, race, and sex; then only age and race; and then just age. This approach also takes advantage of the additivity of totals in this example. However, each unique combination of dimensions defines its own cube.
8. Tables and cubes need to be adequately distinguished, and their dissimilar uses understood. A cube is a model for dimensional data. A table is a format for displaying cube data.
9. As a modeling issue, we need to decide whether a cube can exist without values of the measure in each of the cells. We will encounter cubes where data are suppressed for confidentiality reasons, but that is not what is meant here. Suppressed cells still have data associated with them. Suppression takes place in the production of tables, not their underlying cubes.