

Multidimensional Table Description

04/14/2003

1. Background

From: Thomas, Wendy L. (2002). The Changing Face of Data Access. Minnesota Population Center, DIG-IT Presentation (February, 13, 2002)

What are Aggregate Data?

- Aggregate data are the result of manipulating microdata by totaling the number of cases meeting specific criteria; by summing microdata variables for specific subpopulations; by listing cases that meet specific criteria.....
- Aggregate data are a result set derived through manipulation which have a specific relationship to other result sets derived during the same process.

Why are they so difficult to describe?

- Difficult to provide an abstract definition of what constitutes aggregate data
- Secondary data set – it is the result of manipulating primary (micro) data
- Frequently stored in spreadsheets (grids)
- It is an n-dimensional structure displayed in a 1-or 2-dimensional format
- Discrete cells are used as sources of 'look-up' information

What more description do they need?

- Logical relationship between cells
- Nature of the relationship
- Additivity
- Location within a physical storage grid
- Description of how they were created
- Directions for deriving from microdata

Criteria for an acceptable model

- Describe the logical structure including: full structure, each dimension, each cell, the relationship between all parts, and how they are created
- Describe the physical structure including: multipage grids, irregular grids, how to link them together, and how to access them
- Provide support for the following functions: looking up a specific cell of information, rearranging, collapsing or subsetting a logical structure

2. Cubes

From: Ryssevik, J. A discussion of the cube-specification in the DDI. FASTER Project paper. <http://www.faster-data.org/Metadata/papers/index.htm>

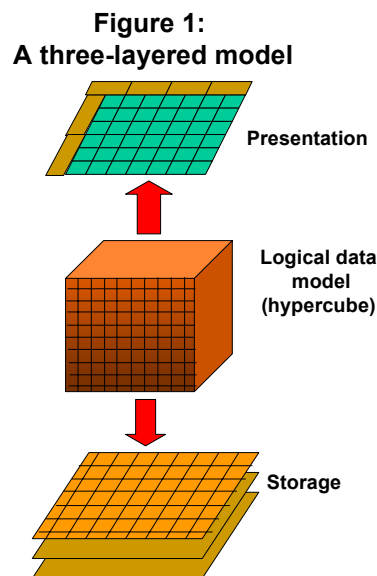
About cubes and aggregated data

Aggregated data will normally be modeled as n-dimensional cubes (hypercubes), where each additional variable constitutes a new dimension. Dimensions will often be hierarchical, including classifications at different levels of aggregation (from detailed to broader groups). Each cell of a cube might include more than one value/measure (based on different methods of aggregation). At the presentational level a multidimensional table will include rules about the position of variables (stub or header), nesting of dimensions etc.

Aggregated tabular data are often produced from micro-data. From one angle aggregated tabular data can therefore be seen as an end product of a statistical analysis more than an input to a statistical analysis. Aggregated tables are also a standard way of presenting and also storing statistical information. In many instances where confidentiality considerations are relevant, aggregated tables are the only data that can be accessed by end users.

The art of Cubism

A major challenge facing any implementation of support for aggregated tabular data is that the data structure is genuinely multidimensional, whereas storage as well as display lives in the confines of a two-dimensional space. Defining the most flexible and efficient mappings or interfaces between the logical data model (the cube) on the one hand and two-dimensional storage and display on the other is the key to this problem (see Figure 1).



Any display of a multidimensional table is inevitably two-dimensional. Even though 3D graphics might give us an illusion of a third dimension, the paper or the computer screen is for all practical purposes two-dimensional. The art of displaying multidimensional data is thus the art of “flattening”. The standard way of flattening a hypercube is of course to display more than one logical data dimension on each physical display dimension, like we do when two or more classifications are nested underneath each-other in the stub or the header of a pivot table.

It goes without saying that any given logical data cube can be flattened in variety of ways, producing a large number of different views of the underlying data. Displayed tables will normally also contain fewer details than the underlying data cube. The table on the screen might display a subset of dimensions or show just the highest and least detailed aggregation level of a hierarchical dimension. It might also focus on a particular “region” of a hypercube, for instance a subset of regional units from a complete geographical classification.

Different ways of flattening dimensions combined with a variety of methods for reducing details, will allow a user to produce an almost limitless number of displayed tables from the same hypercube. Indeed this is exactly what is done when exploring and analysing multidimensional data.

Cf. Additivity

Additivity is a property pertaining to a set of interdependent index numbers related by definition or by accounting constraints. An aggregate is defined as the sum of its components. Additivity requires this identity to be preserved when the values of both an aggregate and its components in some reference period are extrapolated over time using a set of volume index numbers. Although desirable from an accounting viewpoint, additivity is actually a very restrictive property. As already noted, Laspeyres volume indices are additive because extrapolating the base period values by Laspeyres volume indices is equivalent to revaluing quantities in later periods by the same set of base period prices. Additivity implies that, at each level of aggregation, the volume index for an aggregate takes the form of a weighted arithmetic average of the volume indices for its components that uses their base period values as weights. This requirement virtually defines the Laspeyres index. Other volume indices in common use are therefore not additive. (UN 1993 SNA glossary, <http://unstats.un.org/unsd/sna1993/glossform.asp?getitem=10>)

3. DDI nCube (DDI v. 1.3)

A model for describing aggregate/tabular data has been incorporated into Version 1.3 and is now being reviewed and tested by the DDI Committee and other interested parties.

The following example is extracted from the codebook example provided in the DDI site ([Example: Census 1990 Summary Tape File 1](#))

The table is made up to show relationships between variable description, cube description, and data item description (which is for the physical location).

Example: HISPANIC ORIGIN BY RACE

		Hispanic Origin (Var ID="VP10_1") [rank="1"]	
		Not of Hispanic origin (catValue = 1)	Hispanic origin (catValue=2)
Race (var ID="VP6_1") [rank="2"]	White (catValue = 1)	dataltem ID="DI_P10_1" coordNo="1" coordVal="1" coordNo="2" coordVal="1"	dataltem ID="DI_P10_6" coordNo="1" coordVal="2" coordNo="2" coordVal="1"
	Black (catValue = 2)	dataltem ID="DI_P10_2" coordNo="1" coordVal="1" coordNo="2" coordVal="2"	dataltem ID="DI_P10_7" coordNo="1" coordVal="2" coordNo="2" coordVal="2"
	American Indian, Eskimo, or Aleut (catValue = 3)	dataltem ID="DI_P10_3" coordNo="1" coordVal="1" coordNo="2" coordVal="3"	dataltem ID="DI_P10_8" coordNo="1" coordVal="2" coordNo="2" coordVal="3"
	Asian or Pacific Islander (catValue = 4)	dataltem ID="DI_P10_4" coordNo="1" coordVal="1" coordNo="2" coordVal="4"	dataltem ID="DI_P10_9" coordNo="1" coordVal="2" coordNo="2" coordVal="4"
	Other race (catValue = 5)	dataltem ID="DI_P10_5" coordNo="1" coordVal="1" coordNo="2" coordVal="5"	dataltem ID="DI_P10_10" coordNo="1" coordVal="2" coordNo="2" coordVal="5"

* coordNo = rank, coordVal = catValu

nCube Description

```
<nCube ID="NP10" name="P10" cellQty="10" dmnsQty="2">
  <location locMap="LM"/>
  <labl source="producer" level="ncube">HISPANIC ORIGIN BY RACE</labl>
  <universe source="producer" level="ncube" clusion="I">Persons</universe>
  <dmns source="producer" varRef="VP10_1" rank="1"></dmns>
  <dmns source="producer" varRef="VP6_1" rank="2"></dmns>
  <measure source="producer" measUnit="Persons" scale="x1" additivity="y"></measure>
</nCube>
```

Variable Description

```
<var ID="VP6_1" name="P6_1">
  <labl source ="producer" level="var">RACE</labl>
  <catgry ID="CP6_1_1">
    <catValu ID="CVP6_1_1">1</catValu>
    <labl source="producer" level="catgry">White</labl>
  </catgry>
  <catgry ID="CP6_1_2">
    <catValu ID="CVP6_1_2">2</catValu>
    <labl source="producer" level="catgry">Black</labl>
  </catgry>
  <catgry ID="CP6_1_3">
    <catValu ID="CVP6_1_3">3</catValu>
    <labl source="producer" level="catgry">American Indian, Eskimo, or Aleut</labl>
  </catgry>
  <catgry ID="CP6_1_4">
    <catValu ID="CVP6_1_4">4</catValu>
    <labl source="producer" level="catgry">Asian or Pacific Islander</labl>
  </catgry>
  <catgry ID="CP6_1_5">
    <catValu ID="CVP6_1_5">5</catValu>
    <labl source="producer" level="catgry">Other race</labl>
  </catgry>
</var>
```

```
<var ID="VP10_1" name="P10_1">
  <labl source ="producer" level="var">HISPANIC ORIGIN</labl>
  <catgry ID="CP10_1_1">
    <catValu ID="CVP10_1_1">1</catValu>
    <labl source="producer" level="catgry">Not of Hispanic origin</labl>
  </catgry>
  <catgry ID="CP10_1_2">
    <catValu ID="CVP10_1_2">2</catValu>
    <labl source="producer" level="catgry">Hispanic origin</labl>
  </catgry>
</var>
```

DataItem Description

```
<dataltem ID="DI_P10_1" source="producer" nCubeRef="NP10">
  <CubeCoord ID="CC_P10_1_1" source="producer" coordNo="1" coordVal="1"/>
  <CubeCoord ID="CC_P10_1_2" source="producer" coordNo="2" coordVal="1"/>
  <physLoc source="producer" recRef="REC_1" startPos="706" width="9" endPos="714"/>
</dataltem>
<dataltem ID="DI_P10_2" source="producer" nCubeRef="NP10">
  <CubeCoord ID="CC_P10_2_1" source="producer" coordNo="1" coordVal="1"/>
  <CubeCoord ID="CC_P10_2_2" source="producer" coordNo="2" coordVal="2"/>
  <physLoc source="producer" recRef="REC_1" startPos="715" width="9" endPos="723"/>
</dataltem>
<dataltem ID="DI_P10_3" source="producer" nCubeRef="NP10">
  <CubeCoord ID="CC_P10_3_1" source="producer" coordNo="1" coordVal="1"/>
  <CubeCoord ID="CC_P10_3_2" source="producer" coordNo="2" coordVal="3"/>
  <physLoc source="producer" recRef="REC_1" startPos="724" width="9" endPos="732"/>
</dataltem>
<dataltem ID="DI_P10_4" source="producer" nCubeRef="NP10">
  <CubeCoord ID="CC_P10_4_1" source="producer" coordNo="1" coordVal="1"/>
  <CubeCoord ID="CC_P10_4_2" source="producer" coordNo="2" coordVal="4"/>
  <physLoc source="producer" recRef="REC_1" startPos="733" width="9" endPos="741"/>
</dataltem>
<dataltem ID="DI_P10_5" source="producer" nCubeRef="NP10">
  <CubeCoord ID="CC_P10_5_1" source="producer" coordNo="1" coordVal="1"/>
  <CubeCoord ID="CC_P10_5_2" source="producer" coordNo="2" coordVal="5"/>
  <physLoc source="producer" recRef="REC_1" startPos="742" width="9" endPos="750"/>
</dataltem>
<dataltem ID="DI_P10_6" source="producer" nCubeRef="NP10">
  <CubeCoord ID="CC_P10_6_1" source="producer" coordNo="1" coordVal="2"/>
  <CubeCoord ID="CC_P10_6_2" source="producer" coordNo="2" coordVal="1"/>
  <physLoc source="producer" recRef="REC_1" startPos="751" width="9" endPos="759"/>
</dataltem>
<dataltem ID="DI_P10_7" source="producer" nCubeRef="NP10">
  <CubeCoord ID="CC_P10_7_1" source="producer" coordNo="1" coordVal="2"/>
  <CubeCoord ID="CC_P10_7_2" source="producer" coordNo="2" coordVal="2"/>
  <physLoc source="producer" recRef="REC_1" startPos="760" width="9" endPos="768"/>
```

```
</dataItem>
<dataItem ID="DI_P10_8" source="producer" nCubeRef="NP10">
  <CubeCoord ID="CC_P10_8_1" source="producer" coordNo="1" coordVal="2"/>
  <CubeCoord ID="CC_P10_8_2" source="producer" coordNo="2" coordVal="3"/>
  <physLoc source="producer" recRef="REC_1" startPos="769" width="9" endPos="777"/>
</dataItem>
<dataItem ID="DI_P10_9" source="producer" nCubeRef="NP10">
  <CubeCoord ID="CC_P10_9_1" source="producer" coordNo="1" coordVal="2"/>
  <CubeCoord ID="CC_P10_9_2" source="producer" coordNo="2" coordVal="4"/>
  <physLoc source="producer" recRef="REC_1" startPos="778" width="9" endPos="786"/>
</dataItem>
<dataItem ID="DI_P10_10" source="producer" nCubeRef="NP10">
  <CubeCoord ID="CC_P10_10_1" source="producer" coordNo="1" coordVal="2"/>
  <CubeCoord ID="CC_P10_10_2" source="producer" coordNo="2" coordVal="5"/>
  <physLoc source="producer" recRef="REC_1" startPos="787" width="9" endPos="795"/>
</dataItem>
```