How to Markup Aggregate Data in the DDI

William C. Block
Minnesota Population Center
June 2003

!Draft!

This document represents the beginnings of a document that describes how to mark up aggregate data using the "aggregate extensions" that were recently made part of the DDI. It is not complete, and I am interested in hearing from people who find this document useful or have suggestions for improving it. You can contact me at block@pop.umn.edu.

Definitions

<nCube>: An <nCube> is a mathematical matrix of between one and n dimensions where each and every cell of the matrix intersects each dimension at one and only one point. If the <nCube> is additive, the sum of the cells equals the universe of the <nCube>.

Using correct terminology. The table below is a *one dimensional table* that has three cells. It is not a 1x3 or 3x1 dimensional table. This table does not have three dimensions…it has three cells and only one dimension.

10
7
3

<18	10
18-64	7
65+	3

If this table were additive, the universe would equal 20.

<var> A <var> is used to define each dimension of an <nCube> and defines the points along a dimension using <catgry>. A one dimensional <nCube> has one <var>, a two dimensional <nCube> has two <var>'s, and so on.

<catValu> identifies the coordinate value 1, 2, 3, … n of each point along the <var>. For example, male may be <catval> 1, female <catval> 2, etc.

History of the <nCube>?…why is it called <nCube>? This is the infamous Voorburg Compromise involving some tall Europeans (Jostein Ryssevik, Simon Musgrave, and Emiel Kaper) and some short Americans (Wendy Thomas and Bill Block).

The European contingent didn't want to call this idea a "matrix" because that term already meant something specific in European data circles. We (Wendy and I) didn't want to use the term "cube" because that implies something with only three dimensions and aggregate data can have more than three dimensions. We compromised on the term "nCube" which is a cube of "n" dimensions (or a matrix).

Three steps to marking up aggregate data.
Step 1: Mark up each <var>.
Step 2: Mark up <nCube>
Step 3: Markup the <locMap>

Example 1: Marking up a one-dimensional, 3-cell table.

Table: Special People by Age
Universe: Special People

Age

<18	10
18-64	7
65+	3

Step 1: Mark up each <var>.

```
<var ID="VZ">
  <labl>Age</labl>
  <catgry>
    <catValu>1</catValu>
    <labl><18</labl>
  </catgry>
  <catgry>
    <catValu>2</catValu>
    <labl>18-64</labl>
  </catgry>
  <catgry>
    <catValu>3</catValu>
    <labl>65+</labl>
  </catgry>
</var>
```

Do remaining <var>'s as above (this example only has one var, Age, so go on to step 2).

Step 2: Mark up <nCube>

```
<nCube ID="NX" dmnsQnty="1" cellQnty="3">
  <labl>Special People by Age</labl>
  <universe>Special People</universe>
  <dmns rank="1" varRef="VZ"/>
  <measure measUnit="Persons" scale=”x1” additivity="stock"/>
</nCube>
```

The label for an <nCube> is generally the combination of the variable labels and sometimes the universe statement using the operand "by". (The Universe statement is used when it helps to clarify the content of the <nCube>.)

There will be one <dmns> element for each dimension specified in "dmnsQnty". In other words, the number specified in "dmnsQnty" should always match the number of <dmns> marked up in an <nCube>

<dmns> has two required attributes. They are: "rank" and "varRef". The attribute "rank" tells the ordering of the coordinate order for the cells in an <nCube>. (This isn't as hard as it sounds…for example, in a table "Age by Sex," "Age" is first so it has a rank of 1, and "Sex" is second, so it has a rank of 2). The attribute "varRef" is a pointer that points to the ID of the variable that is being described.

The attribute "cellQnty" is the product of the number of categories in each dimension used in the <nCube>. In the special case of a one dimensional <nCube> "cellQnty" is equivalent to the number of categories used in the <nCube>'s single dimension.

The element "measure" gives information about the measurement features of the cell content. The attribute "measUnit" records the measurement unit -- persons, families, households, dollars, etc. "Scale" records the unit of scale -- Does your cell reflect a decimal? A whole number? Is the number in 100's or 1000's? "Additivity" records type of additivity, such as "stock," "flow," "non-additive."?

What I know so far:

Based on the above at this point you can describe the following about a table: Title, universe statement, and its physical structure, including labels.

The table is a one dimensional table with three cells.

Special People by Age
Universe: Special People

Age	Special People

<18	
18-64	
65+	

You now know everything about the table but the numbers in the cell and where to find them. (You haven't yet had to reference the data, but you know the structure of the table).

Step 3: Mark up the <locMap>.

The purpose of a <locMap> is to link what is known about the description of a data item (aka, a cell in an <nCube>), to the physical location of the cell content.

If you are trying to identify a data item, you need to be able to identify which cell in an <nCube> you are talking about, as well as the physical location of the cell content.

(Note: the assumption at this point is that DDI section 3.1 (the physical file description) is already complete!)

Need one <cubeCoord> for each dimension in an <nCube>.

The element <cubeCoord> tells you in dimension order the coordinate value of a particular cell.

The attribute "coordNo" is taken from the rank attribute of the <dmns> in <nCube>.

The attribute "coordVal" is taken from the <catVal> of the <catgry> of the <var> that describes the cell being described by the data item.

```
<locMap> ID="LM"
  <dataItem ID="A" ncubeRef="NX">
    <cubeCoord coordNo="1" coordVal="1"/>
    <physLoc recRef="Rec1"[1] startpos="1" width="9" endpos="9"/>
  </dataitem>
  <dataItem ID="B" ncubeRef="NX">
    <cubeCoord coordNo="1" coordVal="2"/>
    <physLoc recRef="Rec1"[1] startpos="10" width="9" endpos="18"/>
  </dataitem>

  <dataItem ID="C" ncubeRef="NX">
    <cubeCoord coordNo="1" coordVal="3"/>
```

[1] Drawn from section 3.1

```
    <physLoc recRef="Rec1"[1] startpos="19" width="9" endpos="27"/>
  </dataitem>
</locMap>
```

This is what the data actually looks like in the file:

00000001000000007000000003
dataItemA dataItemB dataItemC

At this point I can fill in the cells into the table.

Special People by Age
Universe: Special People

Age	**Special People**
<18	10
18-64	7
65+	3

How to mark up a *two dimensional, six cell* table.

	Sex	
Age	Male	Female
<18	14	15
18-64	28	25
65+	8	10

Here is the same table structure listing cell coordinates:

	Sex	
Age	Male	Female
<18	1,1	1,2
18-64	2,1	2,2
65+	3,1	3,2

Three steps to marking up an nCube.
Step 1: Mark up each <var>.

```
<var ID="VA">
  <labl>Age</labl>
  <catgry>
    <catValu>1</catValu>
    <labl>less than 18</labl>
  </catgry>
  <catgry>
    <catValu>2</catValu>
    <labl>18-64</labl>
  </catgry>
  <catgry>
    <catValu>3</catValu>
    <labl>65+</labl>
  </catgry>
</var>

<var ID="VS">
  <labl>Sex</labl>
  <catgry>
    <catValu>1</catValu>
    <labl>Male</labl>
  </catgry>
  <catgry>
    <catValu>2</catValu>
    <labl>Female</labl>
  </catgry>
</var>
```

Step 2: Mark up <nCube>

```
<nCube ID="NY" dmnsQnty="2" cellQnty="6">
  <labl>Age by Sex</labl>
  <universe>Persons</universe>
  <dmns rank="1" varRef="VA"/>
  <dmns rank="2" varRef="VS"/>
  <measure measUnit="Persons" additivity="stock" scale=”x1”/>
</nCube>
```

What I know so far:

Age by Sex
Universe: Persons

	Sex	
Age	Male	Female
<18		
18-64		
65+		

Step 3: Mark up the <locMap>.

```
<locMap> ID="LM"
  <dataItem ID="A" ncubeRef="NY">
    <cubeCoord coordNo="1" coordVal="1"/>
    <cubeCoord coordNo="2" coordVal="1"/>
    <physLoc recRef="Rec1"[2] startpos="1" width="9" endpos="9"/>
  </dataitem>
  <dataItem ID="B" ncubeRef="NY">
    <cubeCoord coordNo="1" coordVal="1"/>
    <cubeCoord coordNo="2" coordVal="2"/>
    <physLoc recRef="Rec1"[3] startpos="10" width="9" endpos="18"/>
  </dataitem>
  <dataItem ID="C" ncubeRef="NY">
    <cubeCoord coordNo="1" coordVal="2"/>
    <cubeCoord coordNo="2" coordVal="1"/>
    <physLoc recRef="Rec1"[4] startpos="19” width="9" endpos="27"/>
  </dataitem>
  <dataItem ID="D" ncubeRef="NY">
```

[2] Drawn from section 3.1
[3] Drawn from section 3.1
[4] Drawn from section 3.1

```
    <cubeCoord coordNo="1" coordVal="2"/>
    <cubeCoord coordNo="2" coordVal="2"/>
    <physLoc recRef="Rec1"[5] startpos="28" width="9" endpos="36"/>
  </dataitem>
  <dataItem ID="E" ncubeRef="NY>
    <cubeCoord coordNo="1" coordVal="3"/>
    <cubeCoord coordNo="2" coordVal="1"/>
    <physLoc recRef="Rec1"[6] startpos="37" width="9" endpos="45"/>
  </dataitem>
  <dataItem ID="F" ncubeRef="NY>
    <cubeCoord coordNo="1" coordVal="3"/>
    <cubeCoord coordNo="2" coordVal="2"/>
    <physLoc recRef="Rec1"[7] startpos="46" width="9" endpos="54"/>
  </dataitem>
</locMap>
```

This is what the data actually looks like in the file:

00000001400000001500000002800000002500000000800000010
dataItemA dataItemB dataItemC dataItemD dataItemE dataItemF

Age by Sex
Universe: Persons

	Sex	
Age	Male	Female
<18	14	15
18-64	28	25
65+	8	10

[5] Drawn from section 3.1
[6] Drawn from section 3.1
[7] Drawn from section 3.1

The examples so far have contained tables made up of only one <nCube>. More complicated tables, however, can be comprised of more than one <nCube>.

Let's now do a table with

Percentage of Household Type with or without own children
Universe: Households

Total	100
Family Household	66
with own children	44
without own children	22
Nonfamily Household	34

Percentage of Household Type with or without own children
Universe: Households

Total	100
Family Household	66
with own children	44
without own children	22
Nonfamily Household	34

This table contains three <nCubes>:

	# of dimensions	# of cells (data items)	universe
<nCube> 1	1	1	households
<nCube> 2	1	2	households
<nCube> 3	1	2	family households
		5	

Note that the number of data items (cells) totals 5, the same number of cells in the table, above. Each dataitem is contained in one, and only one, <nCube>.

```
<var ID="VTotal">
  <labl>Total</labl>
  <catgry>
    <catValu>1</catValu>
    <labl>Total</labl>  (While there is a theoretical difference from the above examples, putting “total” in both labels is fine).</catgry>
</var>

<nCube ID="NTotal" dmnsQnty="1" cellQnty="1">
  <labl>Total Households</labl>
  <universe>Households</universe>
  <dmns rank="1" varRef="VTotal"/>
  <measure measUnit="Percent" additivity="stockN"/>
</nCube>

<locMap> ID="LTotal"
  <dataItem ID="A" ncubeRef="NTotal>
    <cubeCoord coordNo="1" coordVal="1"/>
      <physLoc recRef="Rec1"[8] startpos="1" width="9" endpos="9"/>
  </dataitem>
  </locMap>
```

000000100000000066000000044000000022000000034

Now I’m marking up the 2nd <nCube>, with one dimension and two cells.

```
<var ID="VHT">
  <labl>Type of Household</labl>
  <catgry>
    <catValu>1</catValu>
    <labl>Family Household</labl>
  </catgry>
  <catgry>
    <catValu>2</catValu>
    <labl>NonFamily Household</labl>
  </catgry>
</var>

<nCube ID="NHT" dmnsQnty="1" cellQnty="2">
  <labl>Type of Household</labl>
  <universe>Households</universe>
  <dmns rank="1" varRef="VHT"/>
  <measure measUnit="Percent" additivity="stockN"/>
```

[8] Drawn from section 3.1

</nCube>

<locMap> ID="LM"
  <dataItem ID="A" ncubeRef="NHT">
    <cubeCoord coordNo="1" coordVal="1"/>
    <physLoc recRef="Rec1"[9] startpos="10" width="9" endpos="18"/>
  </dataitem>
  <dataItem ID="B" ncubeRef="NHT">
    <cubeCoord coordNo="1" coordVal="2"/>
    <physLoc recRef="Rec1"[1] startpos="37" width="9" endpos="45"/>
  </dataitem>
</locMap>

Now I'm marking up the 3rd <nCube>, with one dimension and two cells.

<var ID="VPC">
  <labl>Presence of own Children</labl> Note that I made a value judgement here…changing with or without children to "presence of children"
  <catgry>
    <catValu>1</catValu>
    <labl>With Own Children</labl>
  </catgry>
  <catgry>
    <catValu>2</catValu>
    <labl>Without Own Children</labl>
  </catgry>
</var>

<nCube ID="NPC" dmnsQnty="1" cellQnty="2">
  <labl> Presence of Own Children in Family Households</labl>
<universe>Family Households</universe>
  <dmns rank="1" varRef="VPC"/>
  <measure measUnit="Percent" additivity="stockN"/>
</nCube>

<locMap> ID="LM"
  <dataItem ID="A" ncubeRef="NPC">
    <cubeCoord coordNo="1" coordVal="1"/>
    <physLoc recRef="Rec1"[10] startpos="19" width="9" endpos="27"/>
  </dataitem>
  <dataItem ID="B" ncubeRef="NPC">
    <cubeCoord coordNo="1" coordVal="2"/>
    <physLoc recRef="Rec1"[1] startpos="28" width="9" endpos="36"/>
  </dataitem>

[9] Drawn from section 3.1
[10] Drawn from section 3.1

</locMap>