DDI Example of Use: [Aggregate Data]

Contributors:

# Business Case

DDI-Views needs to be able to describe “aggregate files” also known as “Ncubes”. In this example, using the Australian Election Study 2013 data, unweighted statistics on the variable G1AGE were computed for all combinations of the variables A4 and STATE (see the SQL query below).







# Data Preparation

The following SAS program was used to prepare the data from the aes\_2013\_01259.csv file:

libname Friday "C:\DDRIVE\projects\various\Dagstuhl\2016\week2\DataDescriptionExamples\DataDescriptionExamples\Friday";

**PROC** **IMPORT** OUT= WORK.AESwhole

 DATAFILE= "C:\DDRIVE\projects\various\Dagstuhl\2016\week2\Da

taDescriptionExamples\DataDescriptionExamples\Friday\aes\_2013\_01259.csv"

 DBMS=CSV REPLACE;

 GETNAMES=YES;

 DATAROW=**2**;

**RUN**;

**data** Friday.aes\_2013\_01259SubsetF;

set work.AESwhole;

keep DivisNum UniqueID DateComp State Division A4 G1Age XG5 Weight PartyABY SwingN;

**run**;

**proc** **sql**;

 create table Friday.AES\_Agregate as

 select A4, State, mean(G1Age) as MeanG1Age, Min(G1Age) as MinG1Age, Max(G1Age) as MaxG1Age, count(\*) as N

 from Friday.aes\_2013\_01259SubsetF

 group by a4, State;

**quit**;

**PROC** **EXPORT** DATA= FRIDAY.Aes\_agregate

 OUTFILE= "C:\DDRIVE\projects\various\Dagstuhl\2016\week2\Dat

aDescriptionExamples\DataDescriptionExamples\Friday\AESaggregate.csv"

 DBMS=CSV REPLACE;

 PUTNAMES=YES;

**RUN**;

# Relevant Objects from the Model

## Minimum Description Set

|  |  |  |
| --- | --- | --- |
| **Item** | **DDI 4 Construct** | **Notes** |
| Variable name | InstanceVariable.name |  |
| Variable label | InstanceVariable.displayLabel |  |
| Variable type | InstanceVariable.hasIntendedDataType |  |
| Variable value format | ValueMapping.physicalDataType |  |
| Variable value range | InstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Broken? Could be modeled as a class which allows for repeating segments within a range. Could change the SubstantiveValueDomain.DescribedValueDomain cardinality to 0..n 0..n. Same for SentinelValueDomain |
| Missing value | InstanceVariable.SentinelValueDomain |  |
| Statistics | Not in the model, other than the StatisticalSummary class in the “Keep” package | Not needed for minimum descriptor |
| Code scheme | InstanceVariable.SubstantiveValueDomain (CodeList, StatisticalClassification, etc.) |  |
| Sequence of variables | CubeLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.ValueMapping | Pairwise ordering of Variables - transitive |
| File name | Need to add – do we have a class representing the physical file? A DataStore is a logical construct (which oh by the way has a character set property – move to a class representing the physical instance?) |  |
| File 1st line variable names | CubeLayout.hasHeader/headerRowCount | CSVW has both, even though somewhat duplicative |
| Delimiter | CubeLayout.delimiter/isDelimited |  |
| File encoding | CubeLayout.encoding |  |
| End-of-line character | CubeLayout.lineTerminator |  |
|  | CubeLayout.hasHeader/headerRowCount |  |
|  | CubeLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable |  |
|  | ViewPoint | ViewPoint contains IdentifierRole MeasureRole, and AttributeRole |
|  | IdentifierRole | Delineates which variables describe dimensions in the cube |
|  | MeasaureRole | Identifies which variables serve as measures in each cell of the cube |
|  | AttributeRole | Identifies which variables, if any, serve as attributes of the Cell. |

# Examples – Object Instances

The resulting CSV file is:

A4,State,MeanG1Age,MinG1Age,MaxG1Age,N

-1,1,16.875,15,18,8

-1,2,12.470588235,-1,19,17

-1,3,13.666666667,-1,20,6

-1,4,10.333333333,-1,17,3

-1,5,16,15,17,4

-1,6,16,15,17,2

1,1,16.085106383,-1,20,470

1,2,16.084142395,-1,20,309

1,3,15.083870968,-1,19,310

1,4,16.0546875,-1,19,128

1,5,15.603773585,-1,23,106

1,6,15.967741935,-1,18,62

1,7,17.3,16,19,10

1,8,16.214285714,-1,18,28

2,1,15.665217391,-1,19,460

2,2,16.012345679,-1,25,405

2,3,14.838283828,-1,19,303

2,4,15.464,-1,21,125

2,5,15.932098765,-1,19,162

2,6,15.3,-1,19,60

2,7,16.7,13,20,10

2,8,15,-1,18,18

3,1,14.864197531,-1,27,243

3,2,15.706806283,-1,22,191

3,3,14.832116788,-1,19,137

3,4,15.328947368,-1,19,76

3,5,14.471910112,-1,19,89

3,6,15.333333333,-1,18,27

3,7,13.1,-1,18,10

3,8,16.6,15,18,10

4,1,14.428571429,-1,18,42

4,2,13.711111111,-1,19,45

4,3,12.289473684,-1,19,38

4,4,12.8,-1,18,15

4,5,13.823529412,-1,19,17

4,6,11.875,-1,18,8

4,7,17,17,17,1

### File-Level Metadata:

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| Need to add – do we have a class representing the physical file? A DataStore is a logical construct (which oh by the way has a character set property – move to a class representing the physical instance?) |  |
| CubeLayout.hasHeader/headerRowCount | hasHeader = “true”/headerRowCount = 1 |
| CubeLayout.delimiter/isDelimited | Delimiter = “,”/isDelimited=”true” |
| CubeLayout.quoteCharacter | Not applicable |
| CubeLayout.escapeCharacter | Not applicable |
| CubeLayout.encoding | ASCII |
| CubeLayout.lineTerminator | \n (line feed) |
| CodeList.contains.CodeItem.contains.Code  | [Not in XML – we have a serious problem here. Model is way too deep. Simplify to agree with the pattern.] |
| CodeList.contains.CodeItem.contains.Code.denotes.Category.descriptiveText.content | [Category should be a specialization of Concept] |
|  | Note: Signifier is abstract and cannot be used directly as a datatype. Fix where needed.] |
| IdentifierRole/InstanceVariable – A4IdentifierRole/InstanceVariable – StateMeasureRole/InstanceVariable - MeanG1AgeMeasureRole/InstanceVariable - MinG1Age MeasureRole/InstanceVariable - MaxG1AgeMeasureRole/InstanceVariable - N |  |

# Relationship to Other Standards/Vocabularies

[Describe any useful relationships with other models/standards/namespaces as appropriate, and at a fine level of detail if appropriate.]

# XML Example

[Provide an example of the DDI 4 XML binding for the appropriate part of the model]

# Adherence to Design Principles

[Look at the [DDI 4 Design Principles](https://ddi-alliance.atlassian.net/wiki/download/attachments/37552132/Design_Principles.pdf?version=1&modificationDate=1466520975681&api=v2) and indicate rationale/discuss each in light of this example.]

#