DDI Example of Use:

Unit Record Data in Hierarchical Fixed Records

Contributors: Daniel Gillman (Bureau of Labor Statistics), Arofan Gregory (Aeon Technologies), Larry Hoyle (University of Kansas, Institute for Policy and Social Research), Knut Wenzig (DIW Berlin)

# Business Case

The U.S. Census Bureau releases Public Use Microdata (PUMS) for each decennial Census. Our example is drawn from the 2000 PUMS file for Kansas.

# Relevant Classes from the Model

In this section we will apply the Data Dictionary view, as it combines both logical and physical description. Metadata which is not related to the description of the data, but instead is related to the study is not covered. A UML diagram is included at the end of this document describing the classes used.

## 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 | RectangularLayout.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 | RectangularLayout.hasHeader/headerRowCount | CSVW has both, even though somewhat duplicative |
| Delimiter | RectangularLayout.delimiter/isDelimited |  |
| File encoding | RectangularLayout.encoding |  |
| End-of-line character | RectangularLayout.lineTerminator |  |
|  | RectangularLayout.hasHeader/headerRowCount |  |
|  | ***Need RectangularLayout/SegmentsPerRecord*** |  |
|  | StructureDescription/overview |  |
|  | RecordRelation/InstanceVariableMapping |  |
|  | InstanceVariableMapping/InstanceVariable |  |
|  | RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable |  |

# Examples – Object Instances

Below is the fixed file containing a set of variables from the 2000 PUMS. It has one record per household and a person record for each member of that household. The first and fourth Households, including associated person records are highlighted.



### Household Records

The sample file contained the first 112 characters of the Household records from the 2000 PUMS file. The Census data dictionary is shown below for those columns. This example reads RECTYPE, SERIALNO, STATE, PUMA1, HWEIGHT, TOTPUMA1, and UNITTYPE for households. Other variables are ignored.



### Person Records

The first 27 characters from the 2000 PUMS file were used for the person records. That portion of the Census Data Dictionary is shown below. This example reads RECTYPE, SERIALNO, PNUM, PWEIGHT, SEX, and AGE from person records. Other variables are ignored.



## File-Level Metadata:

DDI-Views can separately describe the layout of each record type.

Identifying Record Types

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| StructureDescription/overview | A description of how to identify the different PhysicalRecordTypes. Probably, no general machine actionable description is possible. In some cases there may be a variable that indicates record type, but in other cases there may only be a description of a general scheme of how the record types are arranged and how they fit together.If appropriate keys exist in variables the RecordRelation in the logical model may be used to describe the relationship among records. |
| RecordRelation/InstanceVariableMapping |  |
| InstanceVariableMapping/InstanceVariable | Example from below, “SerialNo” variable in each record type is used to link the records: InstanceVariable=SerialNo (household)InstanceVariable=SerialNo (Person) |

## Household record:

|  |  |
| --- | --- |
| **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?) |  |
| RectangularLayout.hasHeader/headerRowCount | hasHeader = “false”/headerRowCount = 0 |
| RectangularLayout.delimiter/isDelimited | isDelimited=”false” |
| RectangularLayout.encoding | ASCII |
| RectangularLayout.lineTerminator | \r\n (line feed) |
| ***Need RectangularLayoutSegmentsPerRecord*** | ***1*** |
| 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.] |

## Person record:

|  |  |
| --- | --- |
| **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?) |  |
| RectangularLayout.hasHeader/headerRowCount | hasHeader = “false”/headerRowCount = 0 |
| RectangularLayout.delimiter/isDelimited | isDelimited=”false” |
| RectangularLayout.encoding | ASCII |
| RectangularLayout.lineTerminator | \r\n (line feed) |
| ***Need RectangularLayoutSegmentsPerRecord*** | ***1*** |
| 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.] |

## Column Order Household

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Value** |
| RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable [Is ValueMapping needed? Direct relationship between PhysicalLayoutOrderedPair and InstanceVariable.] | Source = RecTypeTarget = SerialNo |
| RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable  | Source = SerialNoTarget = State |
| RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable  | Source = StateTarget = Puma1 |
| RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable  | Source = Puma1Target = HWeight |
| RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable  | Source = HWeightTarget = TotPuma1 |
| RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable  | Source = TotPuma1Target = UnitType |

## Column Order Person

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Value** |
| RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable [Is ValueMapping needed? Direct relationship between PhysicalLayoutOrderedPair and InstanceVariable.] | Source = RecTypeTarget = SerialNo |
| RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable  | Source = SerialNoTarget = PNum |
| RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable  | Source = PNumTarget = PWeight |
| RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable  | Source = PWeightTarget = Sex |
| RectangularLayout.PhysicalLayoutOrder.PhysicalLayoutOrderedPair.source/target.ValueMapping.formatsInstanceVariable  | Source = SexTarget = Age |

## The values at the variable level:

InstanceVariable: “RecType” (shared)

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| InstanceVariable.name | RecType |
| InstanceVariable.displayLabel | Record Type |
| InstanceVariable.hasIntendedDataType | Codelist (H=Household) (P=Person) |
| InstanceVariable.SentinelValueDomain | NA |
| Not in the model, other than the StatisticalSummary class in the “Keep” package | NA |
| Substantive Value Domain (CodeList, StatisticalClassification, etc.) | CodeList |

InstanceVariable: “SerialNo” (shared)

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| InstanceVariable.name | SerialNo |
| InstanceVariable.displayLabel | Housing/Group Quarters Unit Serial Number |
| InstanceVariable.hasIntendedDataType | Nominal |
| InstanceVariable.SentinelValueDomain |  |
| Not in the model, other than the StatisticalSummary class in the “Keep” package | NA |
| Substantive Value Domain (CodeList, StatisticalClassification, etc.) |  |

HouseHold Layout: Rectype

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| ValueMapping.physicalDataType | Text |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| ValueMapping.SegmentByText.LineParameter | startLine=1startOffset=1endLine=1endOffset=1 |

Person Layout: RecType

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| ValueMapping.physicalDataType | Text |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| ValueMapping.SegmentByText.LineParameter | startLine=1startOffset=1endLine=1endOffset=1 |

HouseHold Layout: SerialNo

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| ValueMapping.physicalDataType | Numeric Code |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| ValueMapping.SegmentByText.LineParameter | startLine=1startOffset=2endLine=1endOffset=8 |

Person Layout: SerialNo

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| ValueMapping.physicalDataType | Numeric Code |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| ValueMapping.SegmentByText.LineParameter | startLine=1startOffset=2endLine=1endOffset=8 |

HouseHold Layout: State

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| InstanceVariable.name | State |
| InstanceVariable.displayLabel | State Code |
| InstanceVariable.hasIntendedDataType | nominal |
| ValueMapping.physicalDataType | numeric |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| InstanceVariable.SentinelValueDomain | Missing = “ ” |
| Not in the model, other than the StatisticalSummary class in the “Keep” package | [table of summary statistics] |
| Substantive Value Domain  | FIPS code (https://en.wikipedia.org/wiki/Federal\_Information\_Processing\_Standard\_state\_code) |
|  |  |
| ValueMapping.SegmentByText.CharacterOffset | startCharOffset=10characterLength=11 |

HouseHold Layout: PUMA1

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| InstanceVariable.name | Puma1 |
| InstanceVariable.displayLabel | Super-Public Use Microdata Area Code (Super-PUMA) |
| InstanceVariable.hasIntendedDataType | nominal |
| ValueMapping.physicalDataType | numeric |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| InstanceVariable.SentinelValueDomain | Missing = “ ” |
| Not in the model, other than the StatisticalSummary class in the “Keep” package | [table of summary statistics] |
| Substantive Value Domain  | Code from Super-Public Use Microdata Area Code (Super-PUMA) CodeList |
|  |  |
| ValueMapping.SegmentByText.CharacterOffset | startCharOffset=19characterLength=23 |

HouseHold Layout: TotPuma1

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| InstanceVariable.name | TotPuma1 |
| InstanceVariable.displayLabel | Total Area of Super-PUMA |
| InstanceVariable.hasIntendedDataType | ratio |
| ValueMapping.physicalDataType | double |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| InstanceVariable.SentinelValueDomain | Missing = “ ” |
| Not in the model, other than the StatisticalSummary class in the “Keep” package | [table of summary statistics] |
| Substantive Value Domain  | Real positive numbers in Arabic numerals |
| InstanceVariable/untiOfMeasurement | Square meters |
| ValueMapping.SegmentByText.CharacterOffset | startCharOffset=72characterLength=85 |

Household Layout: HWeight

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| InstanceVariable.name | HWeight |
| InstanceVariable.displayLabel | Housing unit weight |
| InstanceVariable.hasIntendedDataType | ratio |
| ValueMapping.physicalDataType | float |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| InstanceVariable.SentinelValueDomain | Missing = “ ” |
| Not in the model, other than the StatisticalSummary class in the “Keep” package | [table of summary statistics] |
| Substantive Value Domain  | Real positive numbers in Arabic numerals |
|  |  |
| ValueMapping.SegmentByText.LineParameter | startLine=1startOffset=102endLine=1endOffset=105 |

Household Layout: UnitType

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| InstanceVariable.name | UnitType |
| InstanceVariable.displayLabel | Type of unit |
| InstanceVariable.hasIntendedDataType | nominal |
| ValueMapping.physicalDataType | numeric |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| InstanceVariable.SentinelValueDomain | Missing = “ ” |
| Not in the model, other than the StatisticalSummary class in the “Keep” package | [table of summary statistics] |
| Substantive Value Domain  | Code from UnitType codelist |
|  |  |
| ValueMapping.SegmentByText.LineParameter | startLine=1startOffset=108endLine=1endOffset=108 |

Person Layout: PNum

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| InstanceVariable.name | PNum |
| InstanceVariable.displayLabel | Person Sequence Number |
| InstanceVariable.hasIntendedDataType | nominal |
| ValueMapping.physicalDataType | numeric |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| InstanceVariable.SentinelValueDomain | Missing = “ ” |
| Not in the model, other than the StatisticalSummary class in the “Keep” package | [table of summary statistics] |
| Substantive Value Domain  | Number expressed in Arabic numerals |
|  |  |
| ValueMapping.SegmentByText.LineParameter | startLine=1startOffset=9endLine=1endOffset=10 |

Person Layout: PWeight

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| InstanceVariable.name | PWeight |
| InstanceVariable.displayLabel | Person Weight |
| InstanceVariable.hasIntendedDataType | ratio |
| ValueMapping.physicalDataType | float |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| InstanceVariable.SentinelValueDomain | Missing = “ ” |
| Not in the model, other than the StatisticalSummary class in the “Keep” package | [table of summary statistics] |
| Substantive Value Domain  | Real positive numbers in Arabic numerals |
|  |  |
| ValueMapping.SegmentByText.LineParameter | startLine=1startOffset=13endLine=1endOffset=16 |

Person Layout: Sex

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| InstanceVariable.name | Sex |
| InstanceVariable.displayLabel | Sex |
| InstanceVariable.hasIntendedDataType | nominal |
| ValueMapping.physicalDataType | numeric |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| InstanceVariable.SentinelValueDomain | Missing = “ ” |
| Not in the model, other than the StatisticalSummary class in the “Keep” package | [table of summary statistics] |
| Substantive Value Domain  | Code 1=Male 2=Female |
|  |  |
| ValueMapping.SegmentByText.LineParameter | startLine=1startOffset=23endLine=1endOffset=23 |

Person Layout: Age

|  |  |
| --- | --- |
| **DDI 4 Construct** | **Values** |
| InstanceVariable.name | Age |
| InstanceVariable.displayLabel | Age |
| InstanceVariable.hasIntendedDataType | interval |
| ValueMapping.physicalDataType | numeric |
| ValueMapping.physicalDataTypeInstanceVariable.takesSubstantiveValuesFrom.DescribedValueDomain.minimumValueInclusive/minimumValueExclusive/maximumValueInclusive/maximumValueExclusive | Not applicable |
| InstanceVariable.SentinelValueDomain | Missing = “ ” |
| Not in the model, other than the StatisticalSummary class in the “Keep” package | [table of summary statistics] |
| Substantive Value Domain  | Code 0 - Under 1 year1..89 - 1 to 89 years 90 - Topcode 90+ - State mean of topcoded values |
|  |  |
| ValueMapping.SegmentByText.LineParameter | startLine=1startOffset=25endLine=1endOffset=26 |

# 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.]

# G UML Diagram

#