Data Description Discussion

2015-10-21,2015-10-22, 2015-10-23 Documenting the work at the Dagstuhl 2015 Sprint

# Participants

David Barraclough, Gary Berg-Cross, Michel Dumontier, Martin Forsberg, Dan Gillman, Alejandra Gonzalez-Beltran, Larry Hoyle, Jon Johnson, Daniella Meeker, Steve McEachern, Ørnulf Risnes, Flavio Rizzolo, Dan Smith, Achim Wackerow

# Goal / question

Clarify what is a DataStore?

Capture the discussions on Data, DataPoint, tabular structure and formulate hypotheses on how these objects will interact

# Documentation & notes

# Conclusion

## Proposed Positions

This section contains statements about Datum and other artifacts of DataDescription that have been uttered and/or concluded with over the last year.

Each statement is annotated with an assessment of its validity and additional comments.

## A DataPoint contains the value associated with one Unit for one InstanceVariable

This is true for a dataset in SPSS, SAS, Stata, etc. In R a dataframe cell can contain a list (although this is just one object). In a Cube one might consider that a given combination of value on the dimensions might yield multiple measures.

GSIM seems to have an inconsistency here. In GSIM a datum is “It is the value which populates a Data Point. ” <http://www1.unece.org/stat/platform/display/GSIMclick/Datum> . It only has properties of a Name and a Description.

Yet a DataPoint has many InstanceVariables classified either as identifiers, measures, or attributes while having only one Datum**A Datum is the intersection between a Unit and an InstanceVariable – true or false?**

If true

* One Datum cannot have multiple measure components (countering GSIM)
	+ The systolic and diastolic blood pressure cannot be expressed in one Datum
* One Datum cannot have attribute components (countering GSIM)
* One Datum cannot refer to multiple Units at a time (e.g. Person and Company); instead a *combined* Unit (e.g. Job) has to be defined as the union(?) of the other Units

.

We struggle with this inconsistency - does a Datum have just one value or several?

## An InstanceVariable has a 1:1-relationship with a UnitType – true or false?

**True** but this is an inconsistency with GSIM. GSIM seems to [say otherwise](http://www1.unece.org/stat/platform/display/GSIMclick/Population) on this since the cardinality is 1..\* from Population to UnitType.

GSIM: An InstanceVariable measures a population which is a specification of 1..\* UnitTypes.

##  A Tabular Structure is set of DataPoints where each Datapoint belongs to one row and one column

We may want to describe more than one type of tabular structure. One would be a simple tabular structure with the strict rule that each column contains only values from one InstanceVariable. Another would be a multidimensional structure possibly allowing multiple InstanceVariables in each cell.

## A Row is an ordered set of DataPoints for one Unit

## A Column is an ordered set of DataPoints for one InstanceVariable.

## Agreed positions

## Non agreed positions

We still have not reached a resolution

## Recommendations

# Still open / further discussion

We began our discussions Monday evening

Datum Discussion 2015-10\_19:19:00

The Data Description group met Monday evening for a discussion led by Ørnulf Risnes in which he presented a model that he and Dan Gillman had refined earlier in the evening. The model deals with a data atom which is a single value and all of its associated information. That information is presented in terms of the questions: who, what, when, where, why, how and Value(W5HV). For an example see Table 1. In the discussion we used the term “DATOM” borrowed from the Cognitect Datomic product.

The information in the data atom is a set of Entity, Attribute, Variable, Transaction, tuples (rows), each addressing one of the W5HV questions.

## The Tuples

* The **Entity** is identified by an entity id. All of the tuples having the same Entity ID belong to the same Datum.
* The **Attribute** is the one of the questions (who etc.).
* The **Variable** contains information addressing the question.
* The **Transaction** is a code which is unique for the creation of the Value. Several Values may have come into being as a result of the same process or process step. An example would be the recording of blood pressure, where a value for systolic and diastolic are taken together. In a time series each additional data atom has a new transaction code.

## The Questions

* **What** references a Variable that describes the Value – Its concept, value domain etc.. When data atoms are arranged into a table this would relate to a column of the table.
* **Who** contains a reference to the Unit associated with the value. When data atoms are arranged into a table this relates to the rows of the table.
* **Value** contains the representation of the value of the data atom.
* **Where** points to spatial information related to the value.
* **Why** relates to the purpose of the value
* **When** points to temporal information related to the value
* **How**  describes the manner in which the value came into being – was it via observation or a transformation of other data?
* **Annotation** The data atom may also have a reference to an annotation.

Table 1 The Data Atom Approach

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Entity | Attribute | Variable | Transaction | Predicates |
| dx (datapoint x) | What (Variable) | Ref(Gender) | 1001 | InstanceVariable |
| dx | Who (Unit) | Ref(Ørnulf) | 1001 | UNIT |
| dx | Value | 1 | 1001 | Process output |
| dx | Where |   |   |   |
| dx | Why |   |   |   |
| dx | When | Point/Interval |   |   |
| dx | How | Ref(Observation/Transformation) |   | Source (instantiation of a process) |
| dx | Annotation | NB:be warned | 2006 |   |

The group noted that the distinctions between macro and micro don’t matter this model. It can handle observations about single entities as well as values which are aggregations for an entity. An example follows. A geographic unit might have an identifying code of 1201. An Individual (Ørnulf) might reside in that region. The region might also have a tax rate of 15%. It might also have a total population of 10,000. The code “1201” might serve as answers to different W5HV questions. Queries would allow relating all of those uses.

The following are related by reference to Unit 1201

Who: Ørnulf

What:zipcode

Value:1201

Who:1201

What:tax-rate

Value: 15%

Who:1201

What:population

Value: 10,000

## Structures

We had a brief discussion about how data atoms might be put into structures (rectangular tables etc.) That discussion would continue on Tuesday 2015-10-20. Some thoughts for that discussion:

A DataPoint is a placeholder for a Value.

Building the record will be a query – generating a Collection with an OrderRelation

A query will populate the structure into which the values are placed.

Is the source the output of a process?

Ørnulf’s example – hours worked

Ørnulf: 6:NSD

Ørnulf:12:SSB

Look at Provenance ontology

Models in other communities for instance – city id from registry

“Semantic science integrated ontology”

In RDF would the value always point to Ørnulf in his example? (see his powerpoint)

 DataAtom1 value 42

DataAtom1 who Ørnulf

DataAtaom1 what AGE

Note: We did not discuss another possible “W” – Whether – relating to access rights for the value.

# Tuesday 2015-10-21

The [morning discussion](https://drive.google.com/open?id=1Ym4qGVMU6ppSB3tIY2IP7YqeRXSqFh0CWJmdBOAbCvk) began with the example above .

On Tuesday evening we discussed an example involving the number of hours worked for a particular organization (see figure below). The three values, the number of hours, the id of the organization, and the Id of the person are a set of data that together provide meaning. There was discussion about the Unit(s) involved for the set of numbers. We concluded that there could be one Unit for all three values - the job held by person 230978 for company N027.



The above example is a use case for the need to group values together (in a Datum). We have had the notion of a “Transaction” that groups together values collected together - the classic example is a blood pressure reading having both systolic and diastolic values. We discussed the possibility of using Collections to group InstanceVariables for this purpose.

We’ve also discussed the need to also attach (optional) attributes - e.g. for the blood pressure: sitting vs standing, type of instrument, etc..

Unit and InstanceVariable, though are necessary to put values into a structure - For a table, Unit determines the row, and InstanceVariable determines the column.

# Wednesday Afternoon

## Participants

Dan Gillman, Larry Hoyle, Steve McEachern, Ørnulf Risnes, Flavio Rizzolo

How do we extend the datum structure model we have to support the new required capabilities?

How do we go from DataStore to representation like a CSV?

We reached agreement on the modeling of the ViewPoint - Datum - DataPoint - DataStructure complex. We renamed “Transaction” to Viewpoint. A Viewpoint is a perspective of the data for a specific function (e.g. capture, analysis, storage, etc.). The entity in Table 1 above would be a Viewpoint. A Viewpoint might be that of the original creation/capture of the datum, or it might be from the perspective of a reuse of the data. A Viewpoint could potentially be used to defined a data structure as well (still under discussion). Roles of the Datum might be different in each case, in one a measure, in another an attribute.

Flavio did the following representations:

Here are the models. The first one is the last one we drew on the whiteboard and has all attribute types in the same class:



 In the second one, I made extensions of the Attribute class so that we can manage the vocabularies separately (capture, analysis, etc.)

 

After a discussion, we determined we needed a Record and a Data Structure to be able to create CSV files, for instance. Here is the updated model:



 Just as a reference, this is what we started with a few months ago:

 

# Thursday Morning

See the diagram above.

A Viewpoint is an assignment of the Roles Identifier, Attribute, and Measure to DataPoints in a Record. In a ViewPoint each DataPoint has a single Role. This seems to correspond to a Data Structure Definition in SDMX.

An Identifier is a Role of a Datapoint used to identify a Record under some ViewPoint.

An Attribute is a Role of a Datapoint describing other characteristics of a Record under some ViewPoint.

A Measure is a Role of a Datapoint containing a Datum of interest under some ViewPoint.

A DataPoint is a container for a Datum.

A Datum is the designation of a value .

A Value is a concept with a notion of equality defined.

A Record is a Collection of DataPoints with an optional OrderRelation.

A DataStructure is a Collection of Records with an optional OrderRelation.

A DataStore is a Collection of DataStructures.

# Thursday Afternoon

Building a cube David’s example Final Consumption Expenditures of Households

Dimensions are “year” and “transaction” (Alcoholc Consumption), Measures are Prices in different currencies, Obs\_status are status codes on the cells

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Identifier | Identifier | Measure | Measure | Attribute | Attribute |
| Year | Transaction | Current Price Australian Dollars | Current Price Euros | O bs\_Status1 | Obs\_Status2 |
| 2000 | Alcoholic Consumption | 635 | 411 | E | B |

Question: can we rely on mime types for physical descriptions? - CSV, Fixed width ASCII format.

Final version (?)



## Friday AM

Does Datum need to be a class? Our model sees DataPoints as bins in a Record, with Roles assigned to each bin by a ViewPoint. A Datum can have different roles assigned to it by different ViewPoints grouped together with different Datums. If a Datum is identifiable then its use in different records can be tracked as well as finding its roles in that record.

