Datum Discussion 2015-10\_19:1900

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 The Data Atom

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Entity | Attribute | Variable | Transaction | Predicates |
| dx (datum x) | What (Variable) | Ref(Gender) | 1001 | InstanceVariable |
| dx | Who (Unit) | Ref(Ørnulf) | 1001 | UNIT |
| dx | Value | 1 | 1001 | Process output |
| . | Where |  |  |  |
| . | Why |  |  |  |
| . | When | Point/Interval |  |  |
|  | 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

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

Whiteboard images follow.



