

Issue 18: What is Data Point's Point?

Aside from offering a catchy title, there seems to be some discrepancies in the definition of data point:

Current Situation:

The current version of GSIM has the following definitions:

DataPoint: A Data Point is a placeholder (for example, an empty cell in a table) in a Data Set for a Datum.

UnitDataPoint: A placeholder in a Unit Data Record to contain the value (Datum) for an Instance Variable with respect to a given Unit

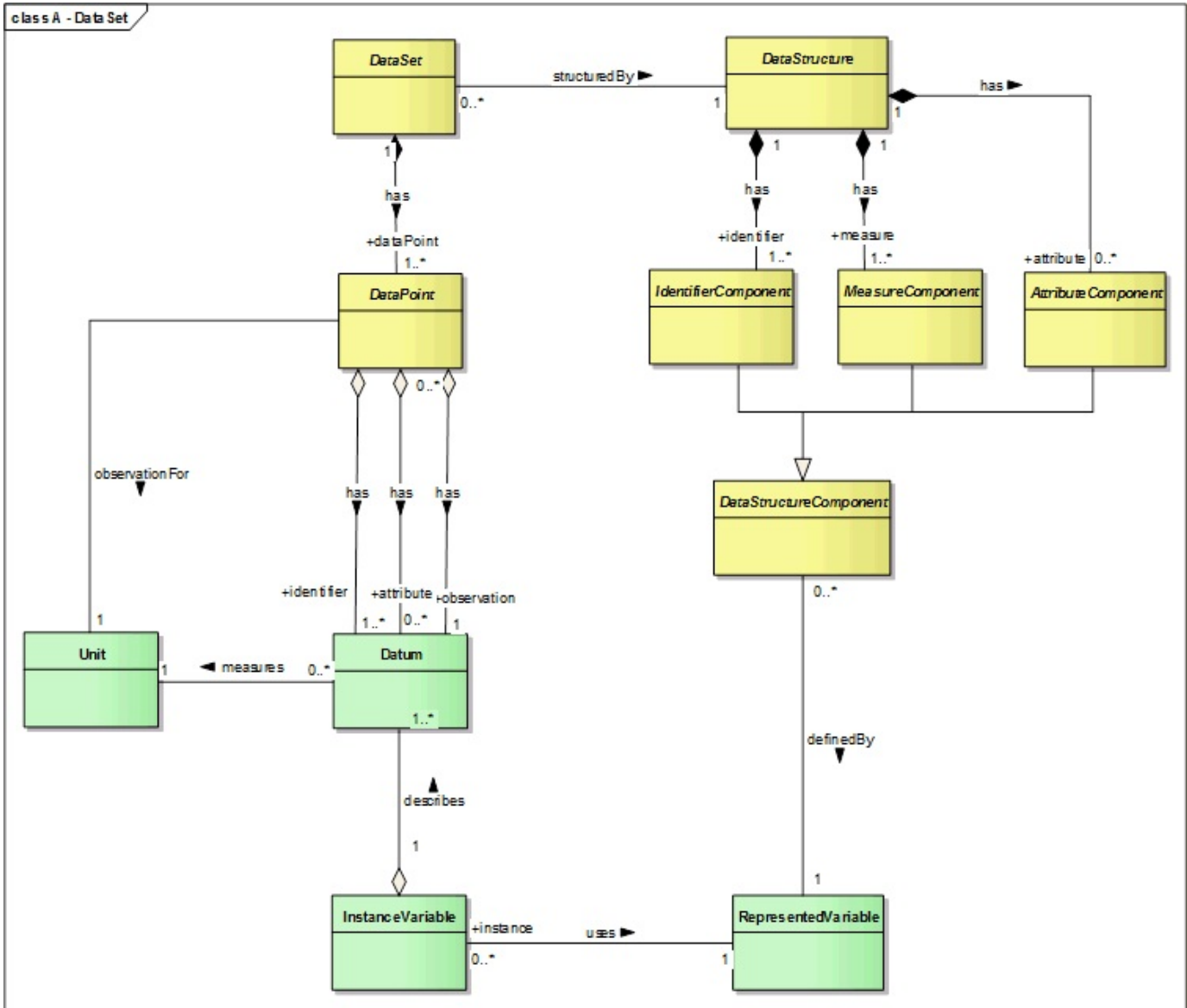
Example: (1212123, 43) could be the age in years on the 1st of January 2012 of a person (Unit) with the social security number 1212123. The social security number is an identifying variable for the person whereas the age, in this example, is a variable measured on the 1st of January 2012. The value can be obtained directly from the Unit or indirectly via a process of some kind.

Issue:

If we consider the above definitions of DataPoint and UnitDataPoint, then they say that it is a cell within a UnitDataSet. The example given within UnitDataPoint, and the corresponding class diagrams (ie fig 44, page 248) however, tend to indicate that DataPoint (just referring to the one for the rest of the discussion) is in fact a combination of information, which would be coming from multiple cells of data (ie, a 'row' of data).

If we go by the written definitions, that the DataPoint is a cell of data, then the example should show that it is the container to hold age in years (43, as per example above), not the combination of social security number and age. There would be another DataPoint which would hold the social security number (1212123) for that Unit.

Within the diagram, DataPoint has Datum (value) related to it in three ways - as an identifier, an attribute, or an observation (although this should be measure, based on the name of the equivalent Data Structure Component). According to the cardinalities in the picture for these three relationships, a DataPoint MUST have at least one identifier, MUST have at least one measure/observation, and may have one or more attributes. This seems to give an indication that the DataPoint covers more than one 'cell' of data. The example also indicates that it is a combination of information - as the DataPoint includes both the age and the social security number for the unit.



Suggested Solution:

Proposal 1: We assume that the definitions of DataPoint, UnitDataPoint and Datum are correct, which means that the cardinalities between DataPoint and Datum are incorrect in Fig. 44?

Therefore, a DataPoint represents a single cell of data, and so the example has two DataPoints - one for age (43) and one for social security number (1212123).

or

Proposal 2: We assume that the definitions are incorrect, and the cardinalities in fig. 44 are correct. Therefore, a DataPoint represents all the data for a single Unit (ie, it is the whole 'row' of data, including both the age (43) and the social security number (1212123) for the Unit.

Other Considerations:

1. Is the information shown in fig. 44 only correct for dimensional data, and so is wrong in the case of unit data?
2. A diagram showing the relationships and cardinalities between UnitDataPoint, UnitDataSet, UnitDataRecord and Datum would be helpful in clarifying things (ie, a combination of the information that is currently split between fig 44 and fig 45 in GSIM v1.0 (pages 240 and 250 respectively).