Draft 2019-09-30

### Long data format

The same data as in the wide example can be expressed in a different format called Long as shown in the figure below. This format is often used to express event data. In the Long format each row now contains a Unit Identifier, a variable identifier, and a value DataPoint. In pure form, each row of a long structure contains a DataPoint with the value of interest along with identifiers for a Unit and a column with a code that identifies the variable that associates with the value in the value DataPoint. In the figure below the Value column contains DataPoints with values from more than one variable, Sex, Born, Died, RefArea, and Longevity. Note that there may be many rows for a Unit (like for “Marie”). There can also be columns containing attribute values. The Verified” column below is an attribute that indicates whether the value in the Value column has been verified.



## Long data format (often used to express event data)

The high-end view of the long model is shown below. Each DataPoint in the dataset is based on one of five components. Each component is associated with a RepresentedVariable that can define a column in the tall table. These perform the following functions:

* IdentifierComponent – one of possibly several components that together identify the Unit associated with the measures and attributes. In the example above this is the CaseID column
* MeasureComponent – a measure just like in the wide layout. This allows a hybrid wide-tall layout. There is no such column in the example above if there were the values for the Marie rows would all be the same.
* AttributeComponent – an attribute that annotates the associated measure values. This is the Verified column above.
* VariableDescriptorComponent – an indicator of the InstanceVariable in each associated VariableValueComponent DataPoint. This is the VariableRef column above. In the first row the code “Sex” indicates that the value “Female” is associated with the variable named “Sex” used in the Wide table. Note that this component has an association to a specific VariableValueComponent.
* VariableValueComponent – defines a column that has a value associated with the value in the VariableDescriptorComponent. This is the Value column above. The “3.3.1932” is interpreted as the date that Marie was born. This column will have to have datatype as generic as needed to hold all of the values from the set of variables indicated in the VariableDescriptorComponent. In the example above there is a mix of numeric (Longevity), Date(born, died), character(Sex), and geographic codes(RefArea) variables. A character datatype for the associated RepresentedVariable would be required. In many statistical platforms there are tools to reshape data between wide and long format. Many have restrictions that would force all of the measure values to have the same datatype (e.g. all numeric).



The diagram above hides a little complexity involving the association between the LongKey and the LongKeyMember. The LongKey is actually a composite of LongKeyMembers, each of which is based on one of the five component types. A LongKey could include, for example, two IdentifierComponents, such as Household and personInHousehold.



This figure shows more detail of how the VariableDescriptorComponent identifies the InstanceVariable associated with a value column. In the first row of the example table, “Sex” is a Descriptor stored in the VariableRef column a VariableDescriptorComponent. That VariableDescriptorComponent refers to the VariableValueComponent “Value” column. Through this mechanism the code “Sex” in one DataPoint describes the meaning of the string “Female” through an InstanceVariable that might be named “gender”. This InstanceVariable would provide information about the concept, value domain, population and more.

