Mapping Instruments to the Variable Cascade

This report contains a discussion of the requirements for mapping instruments to the variable cascade. We use an example to illustrate the ideas.

Much of the work coming out of the first week of the Dagstuhl sprints in 2017 regarding instruments and how they map to the variable cascade focused on a single point of connection. This point is the represented variable. We will show that a looser but still precise connection is required.

Some question specifications require less than what a represented variable requires. For instance, the categories making up the response choices for some question might not include codes. Take

Concept – gender

Question – what is your gender?

Response choices –

Female

Male

Other

Refused

Here, codes are not provided, so a represented variable cannot be fully specified.

Now, sometimes the codes include the ones for sentinel values along with the substantive ones. Take

Concept – gender

Question – what is your gender?

Response choices –

<f, female>

<m, male>

<o, other>

<r, refused>

Here the both substantive and sentinel categories have codes provided for them. This over-specifies a represented variable. Represented variables do not include codes for sentinel categories.

Now, the end goal of the specification of variables from an instrument is the establishment of instance variables. These either allow an instrument to record data or describe data already collected. In both cases, the instance variables provide the description of actual data.

Every question that is specified contains both question text and some description of the allowed answers. The question text, in turn, represents some concept. The allowed answers are either a list of categories or a description of allowable categories, such as a number range. Together, the concept and the allowable categories are enough to describe a conceptual variable. This means the minimum specification is that – a conceptual variable. And, we already know the maximum is an instance variable. From the examples, we know points in between conceptual and instance are possible, including represented.

However, in the design of the DDI-4 model describing the intersection between instruments and variables, we do not know precisely what the point of intersection is.

We propose to provide the template for an instance variable and a flag indicating whether the filled out template obtained so far (in the development of an instrument for implementation) reaches the conceptual, represented, or instance levels of the cascade. In each case, that level and the level(s) above are all specified. The final result, ready for describing data either about to be collected or already so, is an instance variable.

The advantages to this approach are

* There is no need to create a new kind of represented variable
* The instrument as specified so far is attached to the cascade at the right level given what we know
* There is just one actual connection point – the instance variable
* The connection between the instrument and the cascade is natural, as there is no need to stretch the definition of represented variable in order to accommodate all levels of specifying questions

The modeling and data description teams need to figure out how to represent this new reality. Our guess is that one new class should be added to the model called Instance Variable Template (or some such name) which includes the flag mentioned about and a link to an instance variable (the being filled out).