Linking Instrument, Observations, Datum and Variables

## Dan Gillman and Jay Greenfield

# Background

Linking Instrument, Observation, Datum and Variables in DDI 4 has turned out to be challenging.

We start with certain “facts”:

* Observation is the result of the process of applying a particular Question in an Instrument to some experimental Unit. *There is no observation until a question is asked.*
* Datum results when one records an observation. This may be done without human involvement and, for that matter, may be done along with the observation. For example, a survey system records an observation. *Importantly, Datum can be re-recorded. This is often referred to as Data Processing.*
* Each recording of a Datum has a data description. This is where information models like ISO 11179 and the GSIM/DDI4 variable cascade come in. *In DDI 4 the Instance Variable (IV) is part of the information model for a Datum*
* The DDI 4 variable cascade helps us keep track of the proliferation of Datums as observations are recorded and Datums are copied.
* There is no inherent relationship between a question and a variable. This relationship is the product of data processing and data description. *Without a codebook of some sort (the data description), questions can’t be associated with variables*

Reviewing the work of others, it turns out that there exists a model that can account for at least some of these “facts”. It is a process model that drives much of the thinking in biomedical research today.



We can liken “instructions” to instruments and “actions” to the execution of instructions or, again, the administration of an instrument with a unit (patient). In this model administrative events can be likened to “unit management” in which we might include statistical sampling. Also in this model “paradata” is included in “observations”.

This is the openEHR data collection lifecycle. It is rooted in cognitive psychology and thinks of the data lifecycle as an “investigation”

# Approach

Following openEHR, DDI4 might also use a process model to link Instrument, Observations, Datum and Variable. What DDI4 adds to process models like openEHR is the variable cascade which can make sense of cycles of investigations, re-recording or, again “sampling” not in the statistical sense but in the sense we make when we tweet and re-tweet or when one artist sample the music of another.

# The Model















# Datum and Context

Again openEHR defines an information model which puts together datum and context. In the openEHR information model context both metadata and paradata are included:



Here there are slots (data points) for data (weight), metadata (protocol/device) and paradata (state/state of dress, state/confounding factors) and, separately, events or, again, successive measurements and time.

openEHR refers to this information model as an archetype. The slots are general but for different archetypes each slot can be specialized so that, for example states (paradata) may vary in detail, depending on the archetype.

In DDI4 our information model for storing a datum is a data store. The data store consists of data points, the data points are inter-connected by a data structure, and each data point holds a datum. The DDI 4 approach supports both rectangular data stores that “mash up” data, metadata, paradata and time as well non-rectangular data stores. An example of a non-rectangular store is the StatsNZ data warehouse which hosts and contains “slots” (data points) for different type of facts in the warehouse data store.