04 Variable Relationships Discussion Notes

- **Flavio: Presentation of Variable Cascade in DDI CDI**
- **Barbara: RDA I-Adopt**
- **Get into the vibe of DDI CDI and understand the Domain language**

The change we are proposing to make to the model is to introduce an association between values in different value domains. This will facilitate processing which allows attributes related to variables in a wide format and a single attribute related to the combined measures in a long format to be consistently associated as you transform back and forth between wide and long formats.

Do we need relationships between data points?

We don’t have mechanisms for transformation from non-tabular formats into tabular formats (e.g. XML or JSON into long/wide). This came up when looking for examples of O&M data.

Working documents

Example tables:
https://docs.google.com/spreadsheets/d/11TWQv4QIK8ucbF7_auMfkRtG1o5o7JF0Gz4CqSdHCmM/edit?usp=sharing

Diagrams (draw.io):
https://drive.google.com/file/d/11oSVjX5pCIzylqSgvCXUY0BrxR9cX3o9/view?usp=sharing

Draft paper:
https://docs.google.com/document/d/1rdtlqp0zXVgvBS1cQmx4KPDfOBHDGzb1daM5k2Aa3xg/edit?usp=sharing

George, Flavio and Kathy’s paper

https://iassistquarterly.com/index.php/iassist/article/view/1051

Barbara: I-ADOPT information

I-ADOPT Recommendations: https://doi.org/10.15497/RDA00071
RDA working group notes:
https://www.rd-alliance.org/group/interoperable-descriptions-observable-property-terminology-wg-i-adopt-wg/wiki/i-adopt

Graph embedding techniques for semantic distance

Extension of the model:
https://www.rd-alliance.org/group/interoperable-descriptions-observable-property-terminology-wg-i-adopt-wg/wiki/proposal-extend

Presentation:
https://docs.google.com/presentation/d/12oQghEdaOm5247QA_b-f41R044_UgxrB/edit?usp=sharing&ouid=116766443894578314008&rtpof=true&sd=true

Complex use case: https://doi.org/10.1007/s10765-006-0096-4

Claus-Peter's differences vocabulary

https://zenodo.org/record/8092028

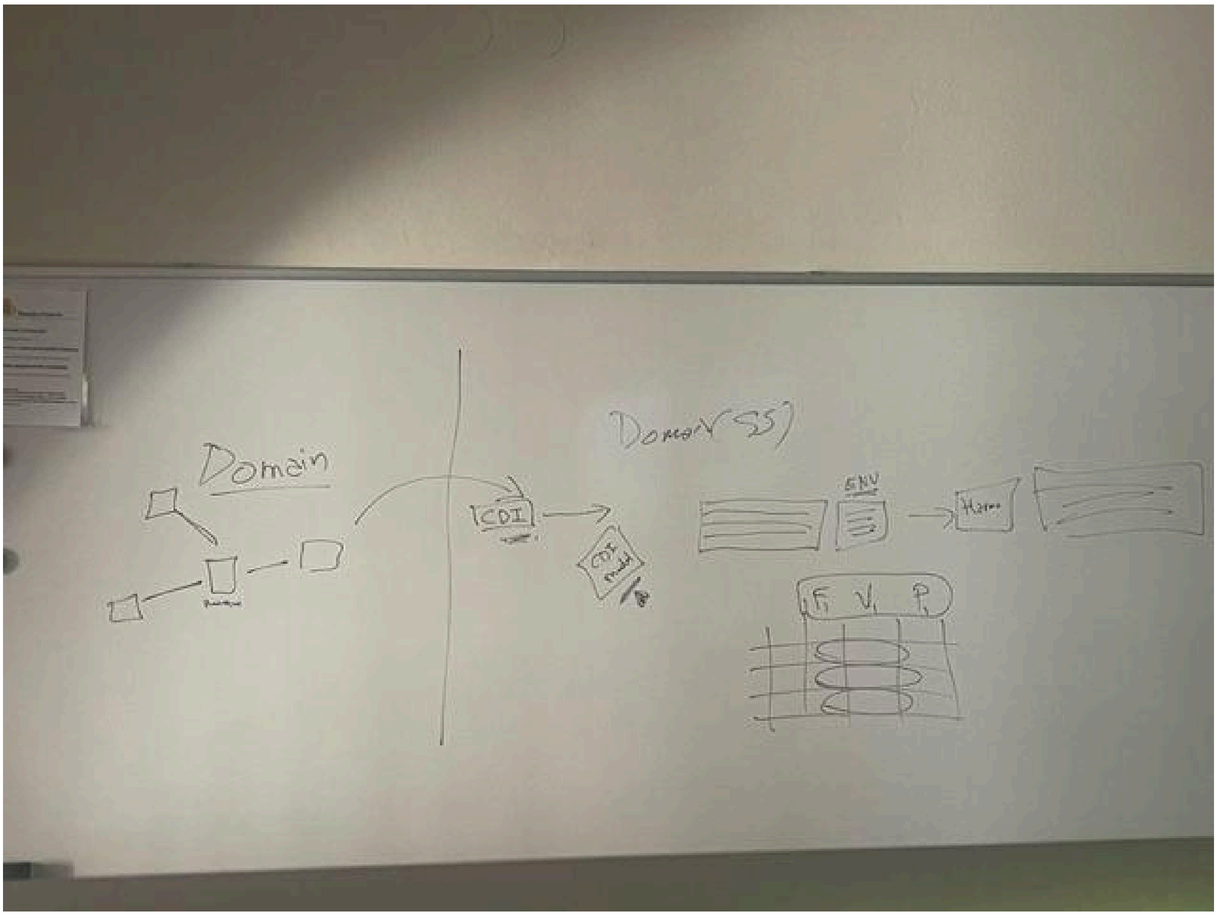
https://vocabularies.cessda.eu/vocabulary/Variables-Relations?lang=en

DDI Tools
Questionnaire Editor: https://multiweb.gesis.org/qeditor/
DDI Search: https://demo.ddi.gesis.org/

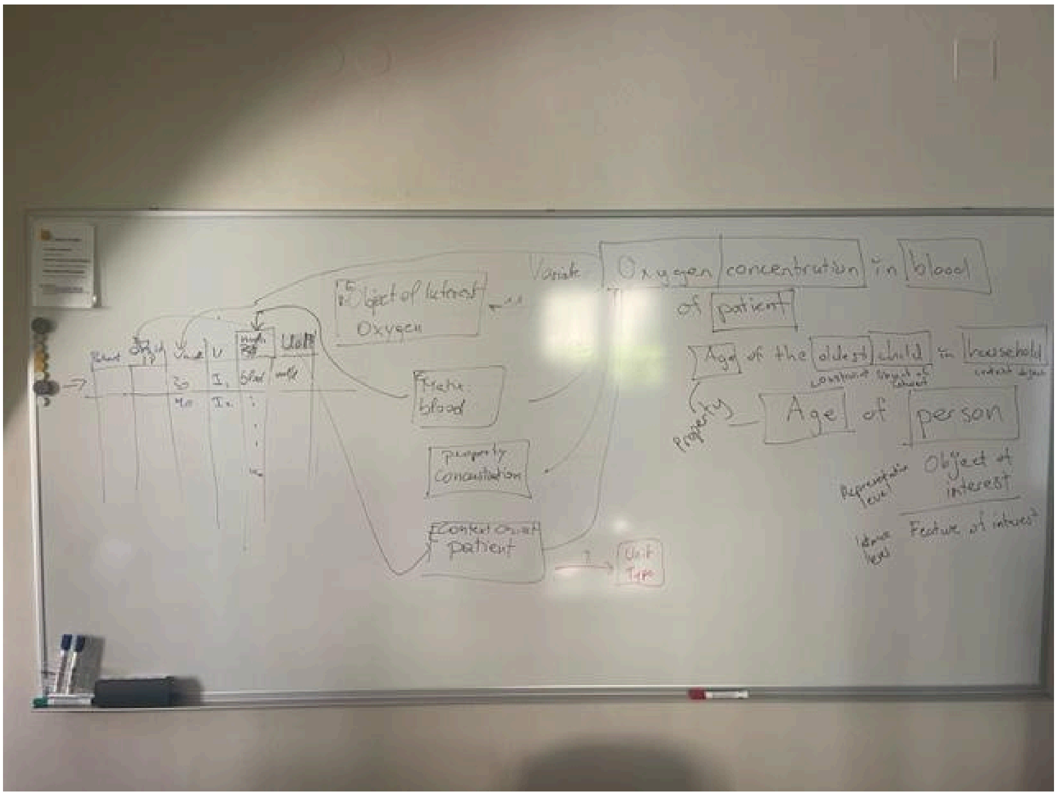
Other references

Tidy data in Python: https://byuidatascience.github.io/python4ds/tidy-data.html

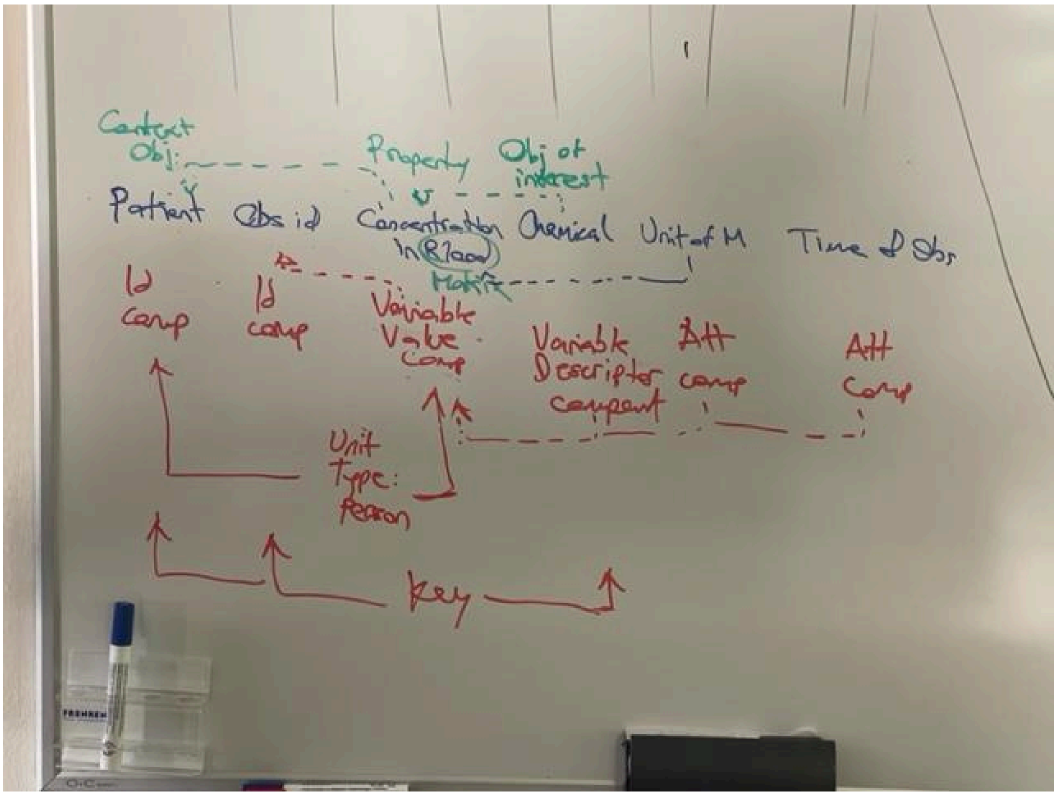
Arofan working example



Mapping I-ADOPT to a data matrix (Tuesday 9am)



Worked example (Tuesday 11am)



SUMMARY (TUESDAY NOON)

We (think we) have mapped I-ADOPT into the CDI structure and we can use the data structures in CDI to map - particularly in long format.

Do we lose anything in the mapping? We might lose the distinctions between ObjectOfInterest, Matrix and Property.

I-ADOPT parses parts of a variable into constituent parts - in CDI we would generally put that into a single definition.

If a domain is described in I-ADOPT, we can move it to CDI. (We have used both I-ADOPT and social science examples).

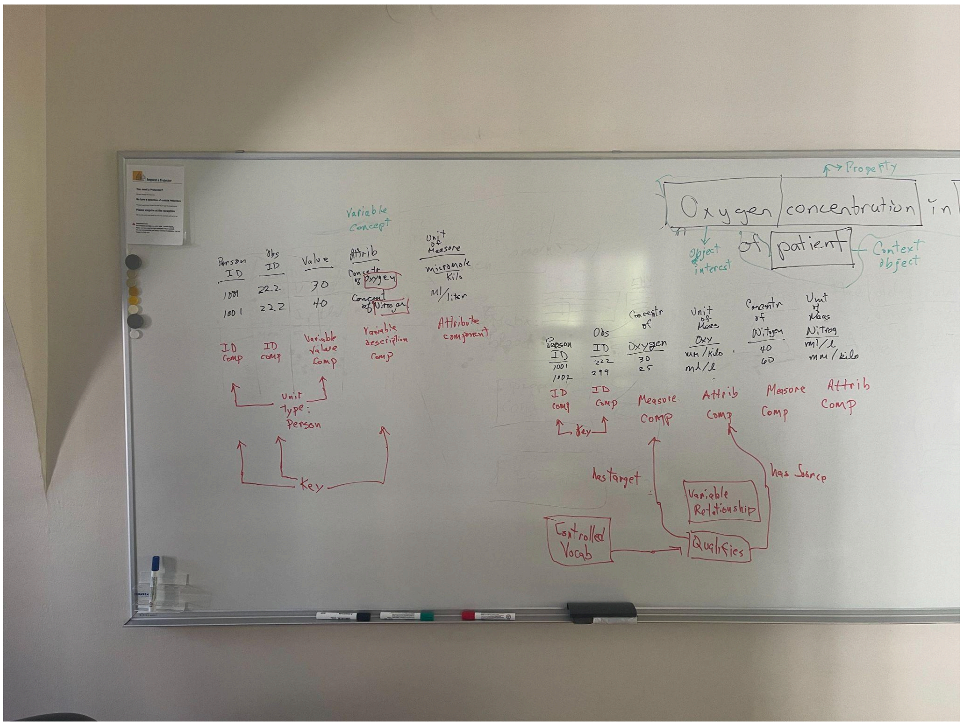
We still have to look at how to represent the dependencies between variables without relying on the data structure.

The “Views on Data and Metadata” paper does reflect many of the use cases we see for I-ADOPT.

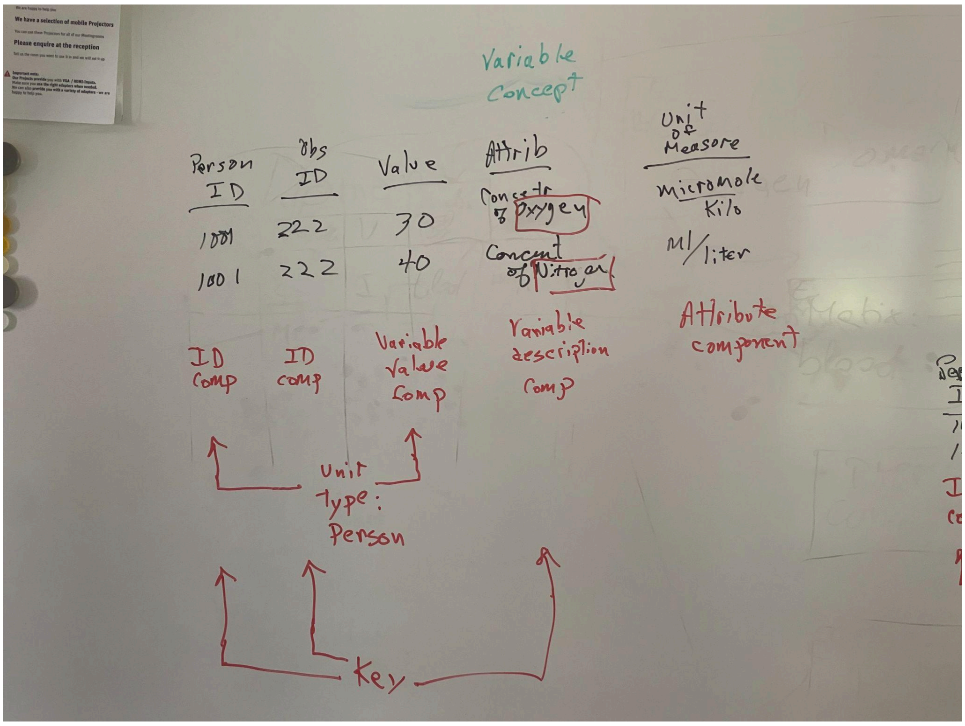
Mapping our Example into I-ADOPT, Long and Wide

We simplified/revised our working example. Following this, we mapped the long and wide formats as follows, to see if we could effectively describe the variable relationships in different formats.

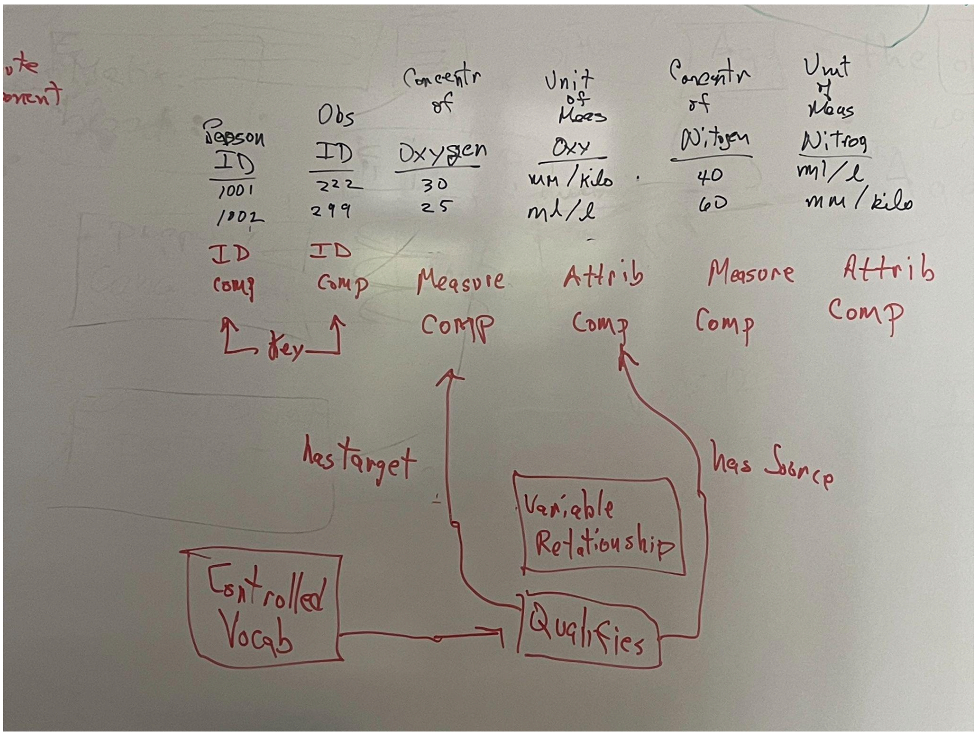
Long and Wide - Mapping CDI objects



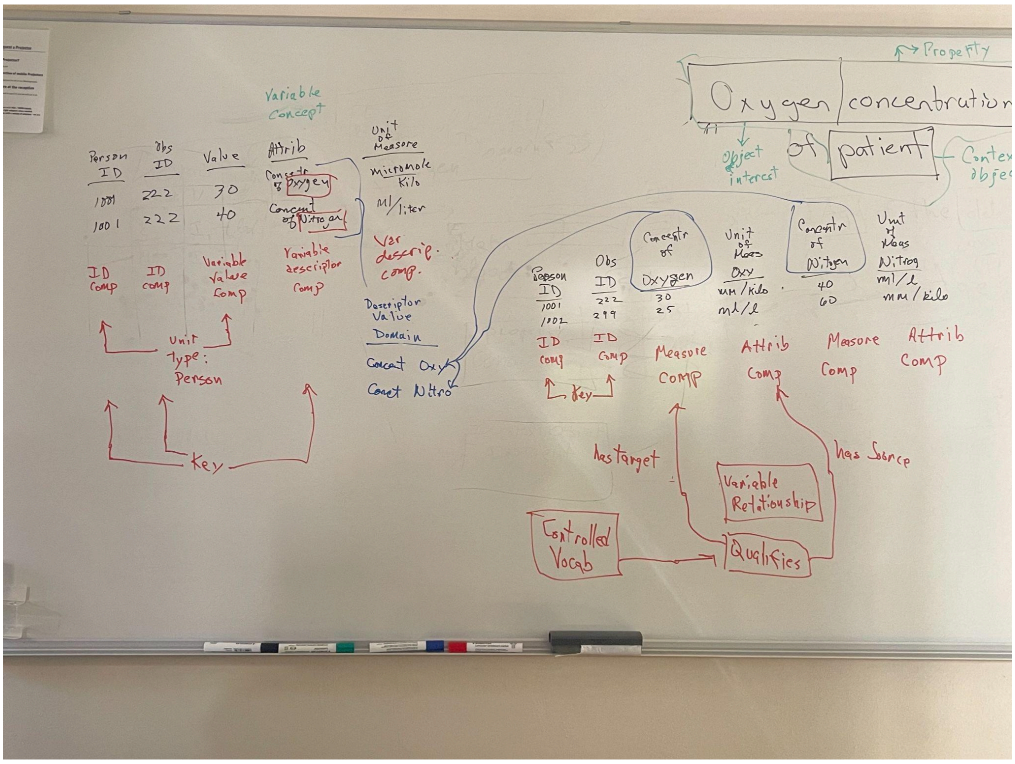
Close-up - Long Mapping (note: Unit of Measure should be VariableDescriptorComponent)



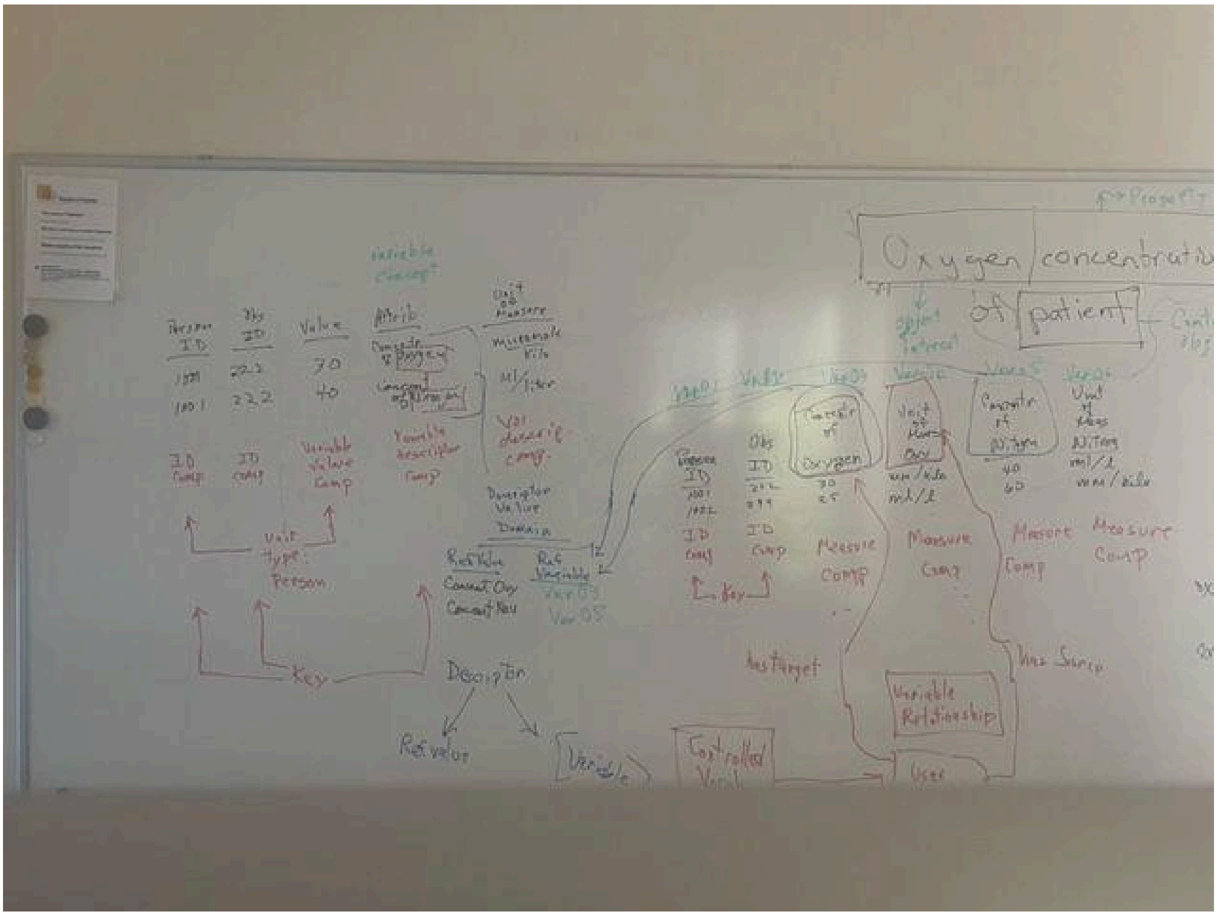
Close-up - Wide mapping (note: Unit of Measure should be VariableDescriptorComponent)



Revision - LongWide mapping



Long-wide mapping - Tuesday 3.15pm



Tomorrow:

Write up the CDI instance of the examples we worked through today

- Provide the instance as a (visualised) graph
- Provide an evaluation of the capacity to represent variable dependencies/relationships
- Provide an evaluation of instance level dependencies
- Provide recommendations for DDI and other standards as to how to go about such representation

Work through an example of an O&M representation to CDI

- Noting that this could be already covered (likely very similar to I-ADOPT)

O&M Example

From the O&M Examples:
http://schemas.opengis.net/om/2.0/

Weather record:
https://schemas.opengis.net/om/2.0/examples/swe_weatherRecord1_t.xml

Weather observation:
https://schemas.opengis.net/om/2.0/examples/weatherObservation.xml

The observation is in XML - we transferred into a (wide) format

Wed PM summary

Long debate about variable relationships

Two proposals:

1. Arofan:

The change we are proposing to make to the model is to introduce an association between values in different value domains. This will facilitate processing which allows attributes related to variables in a wide format and a single attribute related to the combined measures in a long format to be consistently associated as you transform back and forth between wide and long formats.

2. George
LINKED HERE.

Starting to write up our paper
Example tables and diagrams of the CDI instance are now complete.

- Tables here
- Diagrams here

Tomorrow: write up
(Possibly: data point relationships??)

Questions/issues:

Do we need relationships between data points?

We don't have mechanisms for transformation from non-tabular formats into tabular formats (e.g. XML or JSON into long/wide). This came up when looking for examples of O&M data.

Thursday-Friday

Write up of content from discussions, included in the Variable Relationships paper.

Friday morning

There was also a discussion of the possible uses of the I-ADOPT model for official statistics

Notes from discussion (from Franck):

This is a short text summarizing the idea I just exposed. I don’t know if it fits in our report, but I would be happy to discuss it further.

In official statistics, there are often variables with long labels for example "number of nights in a 3-star hotel". The I-Adopt framework can be used here also:

nights -> object of interest
number -> property
hotel -> context object
3-star -> constraint

This kind of decomposition could be useful to derive semantic similarity links between variables and associated contexts. In DDI-CDI, the variable above would be linked to a similar concept ("number of nights in a 3-star hotel"), and this concept would link to a small graph representing the I-Adopt decomposition. Techniques like graph embedding (http://rdf2vec.org) could then be used to evaluate the semantic distance between the initial variables. Note that the I-Adopt decomposition of complex labels could be automated with NLP methods (see Barbara's reference).