DDI/Datum-example 26 January 2015

Author: Ørnulf Risnes, Elin Monstad (NSD)

«Helper» variables/time- and source variables

We frequently see variables appearing in clusters. NSD's data collections from adminstrative registers mostly consist of event-data, where events are recorded for individuals over time. Events include changes in marital status, employment, education level, etc.

The first "helper variable" one notices when dealing with event data in the old "rectangular" form, are start-date and end-date-columns that represent the reference period for the events.

Through the work with RAIRD Information Model (RIM), we found that the start- and end-dates should be treated as datum-level metadata, and not as *data* or variables. Findings in the RAIRD project indicate that treating the dates as datum-level metadata was a good idea, and a way to liberate data from the rectangular format.

This example goes beyond the start/end-variables and discusses something called "source variables".

Source variables are also columns in a rectangular world, and each cell indicates the source of one or more of the substantive variables in the data set. With this example we want to explore whether and how source variables can be liberated from the rectangular format, similar to start/end-dates.

Example: Highest Education Level (Bu_grade)

The variable Bu_grade indicates the highest level of education reached by the individuals in the register. It is one of several similar variables, but this one simply indicates (roughly) how many years of education you have completed.

Catch: The national databases of education information are not necessarily complete, and the information in the administrative registers comes from different sources, some more reliable than others. Researchers need to take the sources into account when interpreting results. In the rectangular format, the source variable Bu_grade_Source is stored in a column as the other variables. Its relationship with Bu_grade is implicit and only reflected in freetext metadata around the sources. You have to know the relationship, and remember to bring the source-variable into the analytical dataset upon ordering data from NSD.

Below is a simplified excerpt of a rectangular representation.

Table 1 - Rectangular file representation of Bu_grade and Bu_grade_Source

person_id	Bu_grade	Bu_grade_start	Bu_grade_end	Bu_grade_Source
00000106	10	1975-06-01		02
00000406	13	1970-11-01	1975-05-31	02
00000406	14	1975-06-01	1977-05-31	04
00000406	10	1977-06-01	1985-05-31	04
00000406	9	1985-06-01	1994-05-31	99
...	...	...	...	...

The **Bu_grade_Source**-variable is coded like this:

01 - 1970 Census
02 - Archive File
03 - Educational Loan Fund
04 - Education Survey 1999
05 - 1980 Census
06 - Education Survey 2011
08 - Education Survey 2012
11 - DUF
31 - Health Personell Authorisation Register
99 - Unknown

Transfer to Datum-based data store

In RAIRD, we don't have any rectangular data[1]. Instead have an "anemic" data store for our register data. As the table below shows, there are only 5 columns at this point. Source is not one of them.

Table 2 - Data Store representation

unit_id	variable	value	startdate	enddate
00000106	BU_grade	10	1975-06-01	
00000406	BU_grade	11	1970-11-01	1975-05-31
00000406	BU_grade	11	1975-06-01	1977-05-31
00000406	BU_grade	12	1977-06-01	1985-05-31
00000406	BU_grade	12	1985-06-01	1994-05-31
00000406	BU_grade	12	1994-06-01	
...	...	...	...	...

In a W5H-perspective, the RAIRD data store currently only supports «who» (unit_id), «what» (variable) and «when» (start-/enddate).

Question 2:

Should we add a 6th column to represent the «how»-perspective (here: source)

Table 3 - Source added to the data store

unit_id	variable	value	startdate	enddate	source
00000106	BU_grade	10	1975-06-01		BU_grade_source
00000406	BU_grade	11	1970-11-01	1975-05-31	BU_grade_source
00000406	BU_grade	11	1975-06-01	1977-05-31	BU_grade_source
00000406	BU_grade	12	1977-06-01	1985-05-31	BU_grade_source
00000406	BU_grade	12	1985-06-01	1994-05-31	BU_grade_source

[1] Instead, researchers compose their own rectangular «working files» by extracting variables from the data store

00000406	BU_grade	12	1994-06-01		BU_grade_source
...	...	...	...	...	...

Alas – we immediately see that the 6th column was not sufficient, because we cannot represent the *value* of the source variable. We need an additional, 7th value column:

Table 4 - 7 columns, both source and source value

unit_id	variable	value	startdate	enddate	source	source_value
00000106	BU_grade	10	1975-06-01		BU_grade_source	02
00000406	BU_grade	11	1970-11-01	1975-05-31	BU_grade_source	02
00000406	BU_grade	11	1975-06-01	1977-05-31	BU_grade_source	04
00000406	BU_grade	12	1977-06-01	1985-05-31	BU_grade_source	04
00000406	BU_grade	12	1985-06-01	1994-05-31	BU_grade_source	99
00000406	BU_grade	12	1994-06-01		BU_grade_source	02
...	...	...	...	...	...	...

Question 2:
But now we need a place to explain the codes for the BU_grade_source-information. Where do we put that?

Thoughts:

Many of the W5H-dimension quickly take on complex lives of their own. The seemingly simple contextual information “electrons” turn out to be fractals instead. Is it possible to pull this off without a graph-like representation?