# Provenance Use Case: Australian Election Study, 2013

Session 3B: Wednesday, October 19, 2016 14:00 – 16:15

*Attendees:* Gary Berg-Cross, Nicholas Car, Arofan Gregory, Larry Hoyle, Gregg Kellogg, Deirdre Lungley, Steve McEachern, Eric Prud’hommeaux, and Michelle Edwards

*Editor*: Michelle Edwards

## ABSTRACT

This session worked through a mapping exercise of a real-world study, the 2013 Australian Election Study to the provenance model. The document captures the graphs created by Nicholas Car and the highlights of the discussion held during the exercise.

### Study abstract:

*Ask Steve to provide little blurb about the survey*

To map any study to the PROV model, one approach could consist of mapping every element in DDI4, however, this may not prove to be very useful. Working from a real-life example and determining which parts are useful and which are not, is the best way to tackle this. Start with a base item, in this case the “Study” and work down to a lower level of granularity. This will help determine what should be included and what items we feel add no added value to include. As we work through the different layers, different linkages and layers become apparent that are not necessarily captured by DDI4 but are captured by PROV. Also note that PROV does not necessarily “care” about all the processes conducted by DDI4. The overall goal is to move from abstract to lower level of granularity, creating a result that will allow readers to understand some aspect of your study at their own comfort level of knowledge.

This exercise started with 6 entities on the board: Study, Data Store, CSV file, SPSS file, DDI XML file, and a codebook; a very high level view of the overall study. By the end of the exercise, 4 major areas or zones of the AES were identified: parent, structure, construction, and forward, and a number of entities, plans, agents, providing a very low level of granularity. Please note that this is not a complete view of the provenance of this study, there are many more aspects that may be mapped to PROV. While working through PROV, always keep in mind what is important and begin from there, do not try to map everything out, as many aspects may not be important. As an example, the Data Description model in DDI4 is static, there is no need to map this to PROV. Remember, provenance has been defined as the “change metadata”, where changes that you are interested in keeping are recorded (see previous paper).

PROV model of the 2013 Australian Election Survey. Created by Nicholas Car



Notes about the model:

* By reviewing the above model, the reader can easily follow the flow of information
* Derived variables are fully modelled in DDI, so there may be no need to model in PROV. Is there any added value to modeling derived variable changes in PROV? For our example, we determined that there was added value, therefore we have a Plan called Variable Generation Algorithm for the prov:Activity called Variable Generation where this information can be captured.
* There are 2 possible perspectives on this model:
	+ The data side of a study – so how is data created and where does it come from – questionnaire -> question -> variables
	+ Structural side of the study – how was the sample collected, stratified, weighting, etc..

Three groups of individuals were identified by the group participants, that would be keenly interested in this type of provenance information:

* Researchers – to determine what has happened to the data they are proposing to use
* Institutions – to gather metrics on use of the data produced at or by their institutions
* Funders – to help determine research impact

The following provenance model depicts the information that can be gathered by these three groups:



To complete this exercise, the above models were also created in TURTLE and listed here.