20161026 - Notes from the Australian RDP IG/RDA DP IG meeting

Attendees:
ANDS: Mingfang Wu, Rowan Brownlee
ANU: Lesley Wyborn
CSIRO: Anusuriya (Anu) Devaraju, John Morrissey, Qing Liu
GA: Laura Sandford, Nick Car
RDA DP IG Chair/The iSchool at Illinois: David Dubin
Ronin Institute: Ruth Duerr

RD-A Data Provenance Interest Group (DP IG) update (Nick & Dave):

- Will work on establishing data provenance working group with a concrete plan. Will present the plan to the next RD-A plenary (April 2017) for endorsement. If got endorsed, the working group will start straight away and last for 18 months. The RD-A DP IG may still keep going for general interest, while DP WG will focus on doing specific things.
- Proposed activities for the DP WG may include:
    - Address use cases from other RD-A WGs. RD-A DP WG is seen as an up group to deal with general provenance issues such as interoperability, a few relevant WGs (e.g. collection attribution group) can then deal with specific implementation from generic approach. It is expected that there will be enough use cases generated by other WGs for the DP WG to take on.
    - Establish mechanism to share use cases. Nick and colleagues at GA will investigate to set up a use case database. The front end may be a Drupal form for other WGs to input their use cases. Capture use cases will be classified and put into the DP WG proposal (with a mix of generic and specific ones).
    - Identify provenance implementation patterns and come out recommendations. Recommendations may include how to get started, what are standards, where are difficult bits, what are required infrastructures.
    - Pingback service: There may be multiple provenance reporting systems set up within an organisation and talking to each other, however there hasn't been seen provenance information is exchanged between organisations. This is a proposal in PROV standard but unimplemented. Nick presented the idea (transforming provenance information between organisations) to the Data Document Initiative group last week and attracted interests. This may be a candidate activity for the RD-A DP WG.
    - Slides Dave&Nick presented in the last plenary are available in two forms: The slides/PDF is on the provenance interest group file section, the original markdown source is available at Github: https://github.com/RDAProvIG/WGplanning/blob/master/P8slides.md.
    - This Github space has been used (by David and RD-A DP IG) and will be used for ongoing notes and progress for the RD-A DP WG. For example, David did a demonstration with the data reproducibility group could be a starting point for demonstrating what it looks like from taking a use case to translate into a provenance vocabulary solution or design pattern.

- Dave and Nick will discuss scope for the WG in next a couple of months. The scope need to be formalised in this December.
-

Future meeting arrangement:

- This Australian DP IG will keep meeting every four weeks. We will keep report on our local activities, but leave a portion of time for RD-A DP WG connection and related activities. Aim to get participants from Europe/Asia as well, so we can all informed each other of ongoing activities.
- Nick will send information about another RD-A DP IG meeting in two weeks' time, for a Europe/US friendly time zone (1:00am-2:00am 8 Nov. AEDT).

Updates for Australian DP IG:

Laura:

- Gave a talk at the eResearch Australasia conference (two weeks ago) about GA's enterprise implementation of provenance management, will give this talk at the Tech Talk next week.
- About to finish human input provenance form, that is part of provenance implementation survey.
- Will test the final form with some people who use manual process to capture provenance information.
- Is looking at provenance reporting within FME to transform into different formats and pipeline process.

Nick:

- Attended the DDI Conference last week. Nick's involvement is to transform the DDI standard/platform to PROV-O. There is an option here to take existing standards/systems to work with PROV-O. For example, it is possible to use the DDI standard to express things as in PROV-O, or subset things from PROV-O. This could be an activity for RD-A DP WG as well.
- RData Tracker is a tool that can be plug into R and track provenance from R programming environment. It is about to add PROV-O export function.
- W3C provenance working group mailing list is still active. People can still sign up to that mailing list if interested.

Ming:

- The Tech Talk next week is about provenance. Two speakers are confirmed: Laura will talk about GA's implementation of provenance, Hamish Holewa will talk about provenance implementation at the Biodiversity and Climate Change Virtual Laboratory. A third speaker is to be confirmed.
- The Australia Software Citation work group discussed future activities for the group. An activity relevant to this group is to align software citation with research reproducibility. The group is also looking at providing use cases for implementing/adoption of the software citation principles as proposed by the Force 11 software citation workgroup. Daniel Katz, Chair of the Force 11 software citation WG, has kindly agreed to join our next group, to introduce future activities from his group.

Rowan:

- First time to this meeting, will be interested in following up activities from this group.

Qing:

- Working on workflow side of provenance. For example, given a workflow provenance, how to help users to compare provenance of different workflow instances, e.g. to track if data are at different granularity.

John:

- CSIRO will generate a lot of use cases. CSIRO is doing a lot different national collections along provenance and characterisation activities.
- CSIRO is also working on governance of vocabs, e.g. vocabs for soil archive. Provenance on building vocabs will be an interesting use case.
- Another area is provenance of sensor data coming from different sources.

Anu:

- Working on consumption side of provenance information. Has developed system to recommend datasets and workflows by using provenance information.
- Also working on attribution of physical samples, e.g. creator and collector of physical samples. Will recommend a minimum set of information for citation.

Lesley:

- NCI is getting 10 TB data from different communities including Satellite, Genomics, and Astronomy etc. Everyone is having different concepts as when and how to versioning datasets. We are looking at this issue of versioning, so when we talk about provenance we know which (version) of data to refer to.

Ruth:

Three use cases from US may be of interest to the RD-A DP IG/WG:

- In last 20 years, NASA has metadata attached within each individual data file from Earth Observation System, metadata include pointers to preceding products, codes and versions. It has provenance information but not in PROV standard (data are pre-date PROV standard). So a question here is how they move forward.
- There is an activity from the Global Change Research Program to track/capture provenance statement from government reports for information e.g. process and
- Work on a data conservancy project at John Hopkins, that is graph based system built on top of Fedorda. The system will track provenance from physical objects such as handling of maps.