BoF on Data Versioning

Lesley Wyborn
National Computational Infrastructure ANU
1. Introductions

2. Why, How and Where of Versioning
   i. Why (Lesley Wyborn)
   ii. How (Jens Klump)
   iii. Where (lightning presentations 5 minutes each)
       • Bob Downs (NASA Socioeconomic Data and Applications Center (SEDAC)
       • Mike Jones (Mendeley)
       • ??????
       • ????
       • New W3C Recommendation (30.8.2016) https://www.w3.org/TR/dwbp/#dataVersioning

3. Do we take this any further?
Why is Versioning Important in Research

• Historical: we have moved from:
  – The traditional mode of a single researcher or individual research team collecting all data used in the research paper,
    • versioning was a relatively straightforward process: there is only one data set and it is unique.
  – Modern team based research where research is based publicly available data sets that are either reused or repurposed to support new directions

• Data are sourced from national/institutional repositories (49 PB in Aus):
  – many of these data sets are being continually added to and/or revised
  – Data are being copied between repositories or to local sites
  – Colocation with HPC or cloud means it is easy to create derived products
Big Data processing in the modern era

- Co-location with petascale computers and data-intensive clouds mean:
  - it is possible to derive seamless data sets from thousands of individual files
  - higher level data products can be generated in very short time frames

- Some of the more mature data centers offer web services access
  - enable users to dynamically select subsets based on user defined spatial and/or time queries
  - rarely are any queries different or even individual researchers the same, particularly where the use graphically draws a spatial bounding box on screen to define the area of interest.
• it is hard for a researcher to cite the exact data extract that was used to support a research project or government investigation
  – Particularly where the source data set is being dynamically modified and/or accessed via dynamic queries.
• With increasing duplication of data sets between the data repositories it is also getting harder to identify the canonical or point of truth data set.
• Agreed procedures for versioning data sets and data products in a systematized way
  – so that it is possible to reference the exact version of the data that was used to underpin the research findings
  – or was used to generate higher level products.

• When do we attach persistent identifiers to data sets
  – Do we associate persistent identifiers with particular versions of data sets
  – if they are constantly changing how do we determine when it should be declared a new version?
  – Who assigns the PiD?
NCI has created a **10 Pbyte** National Environment Research Data Collection

http://geonetwork.nci.org.au.au
Satellite Data Use Case

857,000 Landsat Source Scene Datasets (~52 x 10^{12} Pixels)
I AM THE ONE assigning THE DOI to this data set
1. Introductions

2. Why, How and Where of Versioning
   i. Why (Lesley Wyborn)
   ii. How (Jens Klump)
   iii. Where (lightning presentations 5 minutes each)
       • Bob Downs (NASA Socioeconomic Data and Applications Center (SEDAC)
       • Mike Jones (Mendeley)
       • ??????
       • ?????
       • New W3C Recommendation (30.8.2016) https://www.w3.org/TR/dwbp/#dataVersioning

3. Do we take this any further?