4.49 Provenance of climate data

Bruce Bannerman (Australian Bureau of Meteorology)

▲ Full use case description (click to collapse):

This use case is inspired by one of the conclusions of the UK Parliamentary inquiry into Climate-Gate "*...It is not standard practice in climate science to publish the raw data and the computer code in academic papers. However, climate science is a matter of great importance and the quality of the science should be irreproachable. We therefore consider that climate scientists should take steps to make available all the data that support their work (including raw data) and full methodological workings (including the computer codes...).*" When a climate scientist publishes a paper, he needs to be able to refer reviewers to the source data and software source code that underpins the assertions made within the paper. Climate data are typically time-series and can be quite complex. Data can be sourced from a single National Meteorological and Hydrological Service (NMHS), or from a number of NMHS. Software source code is typically stored within a software revision control repository, such as git.

Climate data may comprise all of the following:

1. A time-series of observations of a specific phenomenon at a single sensor, including:
    - Estimations of the value of an observed property.
    - A detailed understanding of the conditions under which the observation was made.
    -
    - Any changes that have been made to the observation (e.g. due to Quality Assurance processes).
2. A collection of the time-series observations described at #1, of the same phenomenon, at the same time steps, perhaps in the form of a discrete coverage (time-series).
3. A time-series representing the distribution of values of the collection of time-series observations represented as perhaps a:
    - Continuous two dimensional coverage.
    - Continuous coverages as three dimensional cubes.
    - Continuous coverage as n-dimensional models.
    - An ensemble, comprising a number of models.
    - The outputs of some analytical process as a time-series/coverage/cube/model.

So using the description of climate data above, when a paper is published, the scientist needs to be able to refer viewers to:

- The analytical data which underpin the paper (perhaps the n-dimensional continuous coverage time-series at #3).
- The quality assured observations data that the continuous coverages at #3 were derived from at #2 and #1 (with details as to why each change to the Quality Assured observations de-

- The quality assured observations data that the continuous coverages at #3 were derived from at #2 and #1 (with details as to why each change to the Quality Assured observations described at #1 were made).
- The 'raw' observations data described at #1, with details as to the conditions, sensors, etc., that the observation was made under.
- The version of the software source code that was used to manipulate the data at #1, #2, and #3.

This is not a trivial data management problem to address, however its resolution will provide a solid data management grounding for future climate science and help address much spurious debate. Parts of the puzzle are currently being worked on, e.g.:

- The World Meteorological Organization (WMO) has published WMO No. 1131: Climate Data Management Systems Specifications that provide a high level conceptual architecture to address much of the data management issues described above.
- WMO and OGC have developed and are developing relevant logical data models ISO 19156 Observations and Measurements, OGC Timeseries, and WMO METCE. The last two are based on Observations and Measurements.
- WMO has released a standard for describing Observations Metadata, called WIGOS Metadata.
- WMO is about to start work on a logical data model based on Observations and Measurements and Timeseries for describing WMO Observations. A future iteration of this model will need to cater for data provenance.
- W3C has developed PROV-O, which has considerable potential for describing data provenance.

The missing part is: How can the provenance of the collection of climate data and the software used to manipulate it be best modeled and in the future, found via the Internet? The resolution of this issue will be of relevance to many domains.

2.2 Spatial Data on the Web Best Practices, 2.4 Semantic Sensor Network Vocabulary, 2.5 Coverage in Linked Data

5.13 Discoverability, 5.33 Observation aggregations, 5.34 Observed property in coverage, 5.35 Provenance, 5.36 Quality per sample, 5.39 Sensor metadata, 5.40 Sensing procedure, 5.42 Spatial metadata, 5.48 SSN usage examples, 5.51 Support for 3D, 5.56 Time series, 5.62 Virtual observations