Points of note from ACRID project meeting (14-Dec-2010)

Author: Jeremy Tandy

Attendees:

- Andrew Woolf, Bryan Lawrence, Ag Stephens, Sarah Callaghan, Arif Shaon (STFC / BADC)
- Tim Osborn, Colin Harpham (UEA)
- Jeremy Tandy (Met Office)

Andrew Woolf is due to leave STFC in January 2011 to move to the Australian Bureau of Meteorology. In response to this, Sarah Callaghan will take over as project manager.

Please find attached the set of slides presented by Andrew Woolf et al. These include notional project plans and descriptions of work items.

The intention of these notes is to describe the key concerns for the Met Office as a result of the ACRID project …

- The Met Office Hadley Centre (MOHC) is seen by all participants of the project as the provider of the OPERATIONAL version of CRUTEM3; using version control terminology, this is the 'trunk'.
- The primary objective of the ACRID project is to help UEA improve the mechanisms they use to manage the DEVELOPMENT 'branch' of CRUTEM3.
- The processes developed to support UEA will need to be necessarily light-weight in order to operate in a research environment.
- The MOHC's desire to validate the proposals developed within ACRID in the MOHC operational context was raised at the project meeting.
- There is no obligation on the MOHC to adopt the processes proposed within ACRID. The MOHC may CHOOSE to adopt some or all of the processes developed for UEA if they appear to provide utility in the MOHC operational context at the right cost.
- The processes developed to support UEA are considered by ACRID project participants to be representative of those facing the Surface Temperatures Databank project. The ACRID project can be used as a focus to develop an initial proposal for the surface Temperatures Databank.
- The ACRID project proposes the adoption of DOI (Digital Object Identifier, doi.org) as the identifier for the dataset as a whole – e.g. the CRUTEM3 dataset would be assigned a DOI.
- The ACRID project proposes to describe the collection of digital objects that comprise each dataset using a combination of existing standards such as OAI-ORE and the Open Provenance model. The collection of digital objects will be expressed as a 'graph' of *linked data* objects; i.e. a set of RDF triples that one can browse to discover information about the composition of the cited dataset.
- Whilst there is clearly value in exposing the linked data graph for audit purposes, more value will be realised by enabling consumers of these datasets (especially scientists) to discover and browse information presented by the data provider that describes explicitly how the dataset was constructed (i.e. metadata). The view of meeting participants was that this

information, where known, often gets disassociated from the dataset over time. The proposed ACRID approach will ensure that the linkage between metadata and dataset is explicit and web-navigable.

- UEA plan to produce a major update of CRUTEM at some point in the future. This will be known as CRUTEM4. UEA indicated the expectation that the MOHC will continue to serve an operational version of CRUTEM4 incorporating monthly updates. Creation of the operational CRUTEM4 dataset will require *merging* UEA's development branch of CRUTEM with the MOHC's operational version. UEA aim to use the CRUTEM linked data graph to assist in this merge activity.
- A suggestion was made that to support future evolution of CRUTEM, both **UEA and the MOHC may benefit from using a single version control platform** to manage the digital assets of the operational 'trunk' (as published by the MOHC) and UEA's development branches. Such a change is likely to require review of operating procedures at both UEA and the MOHC and is considered (*at least by JT!*) to be **outside the scope of the ACRID project**. No further discussion occurred on this subject at the meeting. MOHC may choose to assess the value of such proposals after the ACRID project is completed. However, there is no aspiration from any of the ACRID project participants to enforce changes to the MOHC's operating procedures.
- A DOI can only be assigned to an object that is not changing. The ACRID project proposes associating the DOI with the 'baseline' dataset each time a major update is published; e.g. CRUTEM3, CRUTEM4 … etc. UEA will develop the linked data graph that describes this baseline dataset. **There will be no obligation on the MOHC to update the linked data graph with respect to the monthly updates that are incorporated into the operational version of the dataset**. The MOHC may CHOOSE to do this if there is a requirement from the community to do so – and this can be achieved at acceptable cost to MOHC.
- Proposed Met Office work items:
    1. Validate that the information provided in the CRUTEM linked data graph will assist the merging of UEA's CRUTEM development branch into the MOHC's operational CRUTEM3 dataset to create the operational CRUTEM4 baseline. *Can MOHC use the linked data graph to understand the constituents / provenance of UEA's CRUTEM.dev and evaluate the differences to the CRUTEM3.operational hosted at the MOHC?*
    2. Consider procedures (and evaluate associated resource implications) for MOHC to re-publish the linked data graph for CRUTEM with each monthly update.
    3. Validate the ACRID project proposals developed for CRUTEM by attempting to build a linked data graph that describes HadCET. Issues arising from this prototype application will be evaluated and (potentially) incorporated back into the ACRID project proposals. Given that the outputs of ACRID should also be applicable to the Surface Temperatures Databank, it is important that the MOHC has the opportunity to validate the proposals and ensure that they are fit for purpose. Use of HadCET to validate ACRID proposals gives the MOHC important leverage in influencing the evolution of an emerging standard for the climatology community.
    4. Peer review of technical proposals.
- Work items (1), (2) and (3) would form part of ACRID Workpackage 5 (April 2011 – July 2011). Work items (1) and (2) have clear dependency on UEA's development of linked data graph for CRUTEM. Work item (4) will run continuously throughout the ACRID project.

- Jeremy Tandy will support colleagues from the MOHC in executing work items (1), (2) and (3). Resource implications for these activities will become clear as the technical proposals are developed. Met Office participants will need to liaise with the ACRID project manager (Sarah Callaghan) early in 2011 to establish indicative resource estimates and availability of (technical) support from BADC / STFC.
- Jeremy Tandy will be responsible for work item (4).
- Contribution of expertise from the MOHC to peer review outputs of ACRID Workpackages 2 (Requirements Analysis, Dec 2010 – Jan 2011) and 3 (Research Data Management, Jan 2011 – May 2011) has not been requested – but would be gratefully received. Please refer to presentation for more details of these workpackages.