ESG FEDERATION WORKSHOP: CMIP5 STATUS UPDATE

- The 'product' DRS field will take two values (output | requested). Only requested datasets at QC level 2 can be replicated.
- Where the output dataset is larger than the requested dataset, the dataset(s) will be published as two separate datasets:
    - requested
    - the remainder of the output dataset not contained in the requested dataset. There are no overlapping datasets.
- There will be no QC flags explicitly defined in the THREDDS catalogs. By definition a published dataset is initially at QC Level 1D (data). When Metafor information is added the level advances to 1M (data + metadata).
- QCL2 will be performed at some gateway sites (DKRZ, BADC, ...) and includes human QC. Provisionally, Metafor is providing a space for the QCL2 information to be entered and associated with a dataset.
- Dataset identifiers must be published into a "namespace", i.e. start with a prefix that can be the name of an institution like "badc.", "pcmdi.", or an activity like "cmip5.". Dataset ids for CMIP5 will start with "cmip5." and therefore follow exactly the DRS nomenclature, with "." (**Balaji: would suggest that project.institution be a required namespace block, i.e both are required, above it says either/or**). separating the different DRS fields. ESGF will provide a registry (initially a web page) that lists the reserved namespaces for identifiers.
- Controlled vocabulary for CMIP5 is defined by the DRS specification, and is case sensitive. The publisher will be lenient as far as case sensitivity of DRS-defined fields, in which case it will substitute the (correct) field value into the THREDDS catalog.

- Status of CMIP5 data node implementation:
- PCMDI:
    - CMIP3 data will be republished to provide TDS catalog URLs to the replication manager
    - One CMIP5 subproject will be created for each publication site, with separate data publisher privilege provided for each subproject. Each subproject will be readable by users with read privilege for CMIP5 (research, commercial) but will be writeable only by the specific publication group.
    - Best practices to include:
        - Sites will implement a file system consistent with the DRS convention, with the understanding that:
            - The THREDDS URLs will have the form http://host/thredds/service/thredds_root/activity/...
            - The BADC drslib will support maintenance of the DRS hierarchy
- BADC
    - OpenID relying party filter must work, be made available
    - Shared configuration for CMIP5, also best practices for CMIP5 publication
- Need to resolve how to propagate new gateway certificate
- GFDL:

- ○ Issue on installation: THREDDS reinitialization (resolved)
  - ○ GFDL will publish to the PCMDI gateway
- ● DKRZ:
  - ○ Sample datasets for test publication must be publishable with CMIP5 project, need correct (CMOR2) fields, correct cmip5 project configuration (esg.ini)
- ● ORNL:
  - ○ Two datanodes, one for production, the other for testing
  - ○ Mix of manual configuration and script install.
  - ○ Active dataset publishing: C-LAMP dataset, ARM and AmeriFlux
- ● NCAR:
  - ○ Two datanodes: production, staging/testing
  - ○ Manual configuration due to: remote DB, virtual-host naming.
- ● JPL:
  - ○ Datanode + Gateway installed and working to serve NASA observations as part of CMIP5 archive
  - ○ Security infrastructure upgraded to latest ORP + Security Filters (1.1.8)
- ● Replication test (BADC / PCMDI):
  - ○ BADC and PCMDI will deploy the latest 1.1 gateway snapshot, with replication implemented
  - ○ Initially both sites will create a small (one or two datasets) based on CMIP3 data
  - ○ Both sites will:
    - ■ Publish the sample datasets (different dataset IDs)
    - ■ Share the RDF records between sites
    - ■ Replicate BADC <-> PCMDI to test gateway functionality
  - ○ When the initial test succeeds, ramp up to system scale
- ● Need to test cross-gateway authorization and registration: Users should be redirected to the gateway that supports the relevant attributes (access to the given resource).
- ● Need to verify that file transfers are verifiable / reliable using checksums: source checksum should be computed once and stored; and be accessible by download client.