ESGF Data Publisher

A ~~day~~ week in the life

Getting the data..

* University of New South Wales (Steve Phipps), UNSW gave us both CMIP5 and GeoMIP data to publish. GeoMIP is Geoengineering Model Intercomparision Project built upon the CMIP5 experiment framework.
* Original data has priority over replication, so do this..
* Jeff Painter gets the data from the author and puts it in a scratch directory on gdo2 (/cmip5/data, css01-cmip5, css02-cmip5).

Gdo2 1

* Staged in a “scratch” directory.
* Non-CMIP5 data use a straight move from one directory to another under the same mount point, usually one of these; /css01-cmip5, /css02-cmip5, /cmip5
* Assuming no previous data…
* (if existing non-cmip5, keep or unpublish and move/delete)
* mv /css01-cmip5/scratch/geomip/output/UNSW /css01-cmip5/data/geomip/output/UNSW

Gdo2 2

* For CMIP5 this is different. Use esgcopy_files.
* It opens each file to determine the proper product and metadata in order to copy or move properly.
* *Important:* This script is not an officially supported part of the ESGCET module. It supports construction of a disk-based archive based on the directory_format_for_copy pattern defined in esg.ini. If the files are for replicated datasets (--replica) the directory_format_for_replica pattern is used instead.

Gdo2 3

* But, it fails..must run it through pydebug since messages make no sense.
* pydebug /export/home/drach/gitwork/esg-publisher/src/python/esgcet/scripts/esgcopy_files **--dry-run** -o unsw.txt --verbose --move cmip5 /css01-cmip5/scratch/cmip5/output1/UNSW
* It turned out the data was from a new institute UNSW, which was not defined in the gdo2 esg.ini file for cmip5. So, it contains a hard-coded list of institutions that are support in cmip5!
* To discover this, I ran the debugger and watched it fail when trying to decypher the .nc files for the institute

Gdo2 4

* Other things to check in the /usr/apps/esg/config/esg.ini
* directory_format_for_copy = /**css02-cmip5**/data/cmip5/%(product)s/%(valid_institute)s/%(model)s/%(experiment)s/%(time_frequency)s/%(realm)s/%(cmor_table)s/%(ensemble)s/%(variable)s
* directory_format_for_replica = /**css02-cmip5**/data/cmip5/output1/%(valid_institute)s/%(model)s/%(experiment)s/%(time_frequency)s/%(realm)s/%(cmor_table)s/%(ensemble)s/%(variable)s

Gdo2 5

* Otherwise, esgcopy_files will move/copy your files to a destination on another file system….
* Lesson: When using esgcopy_files always examine the esg.ini file first and verify your parameters.

Where to Next?

* THREDDS has a limitation of 15K datasets before it begins to run like molasses. We are now publishing to pcmdi7 since pcmdi9 is at that limit.
* UPDATE: This may be solved. I just added a configuration parameter to threddsConfig.xml to turn off static catalog caching… so maybe we can now go back to publishing on both pcmdi9 and pcmdi7…
* Lesson: if TDS running slowly double check catalog caching is FALSE in the threddsConfig.xml file

Preparations

* Make sure permissions are set in order to allow thredds to write into any directories you create.
* -bash-4.1$ umask
* 0002 (octal) 0=rwe, 2=rx
* -bash-4.1$ umask -S
* u=rwx,g=rwx,o=rx (symbolically)
* Lesson: unless this is set rw your thredds pub will fail.

GeoMIP

* Their data is similar to CMIP5
* I created a new project handler for generic CMIP5 like MIP projects since they share many of the same parameters.
* esgsetup –handler
* - Project handlers encapsulate the logic about what metadata should be associated with a project, and how it is represented in the DB, directory structures, etc.

Generic MIPS Handler

* I next modified the default project_handler.py to use the CMIP5 handler but removing specifics to CMIP5 etc. We need this handler so publication can pick up all the proper metadata for these GeoMIP files.
* Sudo -s
* python setup.py --verbose install
* Installed /usr/local/uvcdat/lib/python2.7/site-packages/genericCMIP-1.0-py2.7.egg
* Lesson: run this as root otherwise it will fail to write it.

More prep work

* For GeoMIP I started with existing CMIP5 project stanza and modified about ½ dozen parameters.
* Update esg.ini and esgcet_models_table.txt to add new project models, institute, etc.
* Run “esginitialize –c” but make sure esg.ini points to the right esgcet_models_table.txt

But wait!

* On pcmdi7 if it were the first time we reference data on / cmip5-css01/… Check thredds_dataset_roots in the esg.ini file.
* cmip5_data | /cmip5/data
* cmip5_css01_data | /cmip5_css01/data
* cmip5_css02_data | /cmip5_css02/data
* Lesson: If you muck these up the THREDDS links to your (and possibly previously published files) can be broken!

Other checks in the esg.ini

* checksum = md5sum | MD5
* thredds_service_descriptions (OPeNDAP, etc)
* Pointing to the right project handler in the esg.ini?
* Project parameters… to do so must know your data
* Run ncdump –h on a sample file to view attributes.
* Categories - variables per file – datasetid – product ….

Other issues

* Adding new search category facets.
* Add new group & permission to publish new projects. (see /esg/config/esgf_policies_common.xml and esgf_policies_local.xml)
* For new projects or things I'm not certain about I like to do a test publication on pcmdi11 and let the researcher review and okay, usually a very good idea.
* Data files are sometime bad. Scanning/publishing will complain. Notify originator/remove files/proceed?

Map your future

* /etc/esg.env (sets the proper path, etc.)
* If your directories are right use read-directories otherwise use read-files to build a map of files/datasets and extracts the required metadata for DB tables.
* esgscan_directory --read-directories -i ./esg.ini --project geomip -o unsw-geomip-pcmdi7.txt /cmip5_css01/data/geomip/output/UNSW
* Produces dataset-id | file | … |checksum| type * (could do later using esgupdate_metadata)
* Adding an associated document like Disclaimer?
* This could take a while if you have checksums enabled.

Publishing

* myproxy-logon -t 72 -s pcmdi9.llnl.gov -l ganzberger -p 7512 -o $HOME/.globus/certificate-file (SOLR)
* /usr/local/uvcdat/bin/esgpublish -i ./esg.ini --project geomip --map unsw-geomipRD1.txt --thredds --publish --service fileservice
* Can be published in pieces to confirm each step.
* Lesson: unless you get your certificate publishing to solr will fail (also unpublishing) so be careful.

Unpublishing

* esgunpublish -i ./esg.ini -map unsw-geomip-pcmdi7.txt --database-delete
* --database-delete: Delete the associated local database entry for this dataset. By default, the database information is left intact.
* By default esgunpublish deletes all versions of the dataset. To unpublish a specific version n, specify the dataset as dataset_name#n.
* esglist_datasets --all --select version_name -p version=20130605 --no-header -i /esg/config/esgcet/esg.ini cssef |esgunpublish --use-list - --database-delete

More Lessons off the top of my head

* Don’t assume anything, check before/during/after.
* Sometimes providers don’t tell you the truth (update or replacement e.g. KMA epochs all changed).
* Test unpublishing leaving zombie datasets in SOLR but gone everywhere else.
* “SOLR goes in/out to lunch: Bob’s replication notes”
* Unpublish previous non-CMIP5 datasets, previous versions may not be necessary to keep, ask.

More off the top of my head

* TDS not scaling! FIXED?
* DB/TDS Catalog.xml/SOLR gets out of sync?
* Catalog.xml has dataset entry but no DB/THREDDS catalog/SOLR.
* How to move publications from one node to another?
* Publish updates etc. for existing datasets on the same data node (don't pub one ½ of a dataset on one data node and ½ on another).
* This was a simple example, we haven't tackled replication
* If you notice anything that does not seem right, STOP and investigate....