Replication and Versioning Working Team

Progress Update & Future Roadmap

ESGF F2F Workshop,
Washington, DC, December 2016

Stephan Kindermann, Jerome Raciazek, Matt Prichard, Claire Trenham, Jingbo Wang, Ben Evans, Claire Trenham, Katharina Berger, Jeff Painter, Sasha Aims, Joseph Antony, Eli Dart, Chameron Harr, Antony T. Hoang
Replication and Versioning Team Goals: „Create replication tool for moving data from one ESGF center to another; in addition, preserve versioning history of the ESGF

**Synda replication tool related tasks worked on:**
- Task 15.2 Install and test synda installations at DTNs of core sites
- Task 15.4 Integrate and test synda with globus at core sites
- Task 15.5 Define replication policies (reflecting requirements at sites)
- Task 15.7 Test large transfers and work on optimization

**Key obstacles:**
- Site config issues (DN, gridftp, globus, ..)
- Core site DTN/DN infrastructure changes
  - Most site config issues config issues resolved (iwt, tranfer team, ..)
  - Core site replication realated infra stabilized (DKRZ, PCMDI, BADC, ..)

**Accomplishments:**
- Enabled synda based replication tests
  - gridftp endpoint publication, gridftp CA configuration
  - synda extensions
    - e.g. url_replace option, drs path config, globus online support, gridftp configs, docu
  - setup of DTNs at core sites
2016 Missed Milestones (if any)

No missed milestones, yet behind schedule for the „real CMIP6 replication scenario“

• No policy agreements yet – depends on „tier1“ definitions
  (→ also no technical discussions yet on „policy enforcement tools“ )

• „Real“ replication tests in collaboration with climate network working group delayed

Problems:
– Core sites: infrastructure adaptations („CMIP6“ preparation procurements etc. )
– Tier2 sites: gridftp and globus installations not ready, unstable, ..

• No discussions on „replica publication process“ at core sites yet
  e.g. use and test synda processing option for automatic replica publication

  -- initially probably non automatic replica publication procedure at core sites,
       needs some discussions

• Dashboard integration open
2017 Roadmap

Planned Tasks:

• #1: (Non-Technical) definition of (agile) replication process
  We need up to date information on who replicates (or plans to replicate) what from whom as well as available storage resources
  → up to date information on local replication priorities at core sites
    (e.g. synda „selection file“ repo collecting all replication jobs at sites)

• #2: Technical establishment of replication process at core sites
  → synda / DTN production deployments (+ tests)

• #3: steps towards full automation of replica provisioning
  -- automatic replica publication process

• #4: replication and ESGF near processing integration:
  discussions on local caches, cross site aspects
Additional Resources Needed (if any)

• Additional Resources (storage, network, staff) would not speed up the replication process for the next year – better increase the priority of replication related tasks for existing staff

• Storage and network are bottleneck in the longer run, yet by now core bottleneck are site specific optimizations (new DTNs, better network integration, testing .. )

• CDNOT / replication team interactions needs to be established many site configuration aspects ..
Community Feedback from Survey (if applicable)

Versioning aspects worked on

- Versioning support in esgprep tool
  -- retrieve version info from DRS if available
- Agreement: version must be included in CMIP6 directory structure
  -- (works also with „latest“ links)
- Error flag in PID record in case of dataset version updates (addition, deletion, modification of files)
- → errata service
Collaborations Needed (if any)

• Collaboration with other ESGF working teams
  › ICNWG team (thanks Eli !)
  › Publication team (versioning aspects)
  › Data Transfer Working Team (globus related aspects)
  › (Stats and Dashboard working team)