ESGF METADATA & SEARCH WORKING TEAM (ESGF-MSWT): PROGRESS UPDATE & FUTURE ROADMAP

ESGF F2F Workshop,
Monterey, CA, December 2015

Luca Cinquini [1]
[1] California Institute of Technology & NASA Jet Propulsion Laboratory

Copyright 2015 California Institute of Technology. U.S. Government sponsorship acknowledged. JPL Unlimited Release Clearance Number: CL#14-5101
• Work throughout 2015 has been largely dominated by the ESGF security incident, which had two main consequences for the Publishing & Search Services:
  ▸ All ESGF nodes had to be brought offline, and software stack completely re-installed at each node
  ▸ All data will have to be republished

• The ESGF-MSWT took advantage of this unfortunate situation to execute a much needed upgrade of the ESGF Search Services and underlying Solr metadata index
2015 Major Achievements

• **Upgrade from Solr 3.x to Solr 5.x:**
  ‣ Support for atomic metadata updates
  ‣ Much improved support for geospatial searches
  ‣ Better performance, bug fixes
  ‣ Introduction of SolrCloud architecture

• **Infrastructure improvements:**
  ‣ Solr runs within embedded Jetty container (distributed with Solr)
  ‣ Solr is started/stopped with distribution scripts
  ‣ Expose public (aka “slave”) shard on port 80 to avoid firewall issues with port 8983
  ‣ Metadata are still published to “master” shard on port 8984 which needs to be visible only as localhost
2015 Major Achievements

• **Introduction of “local shard”:**
  ‣ Solr index that is not replicated to other nodes
  ‣ Intended for publishing of data collections that are not distributed across nodes and/or are not of federation-wide importance
  ‣ Data can still be downloaded by all users throughout the federation by using the ESGF search service that is co-located with the shard
  ‣ Promotes scalability of distributed searches
  ‣ Promotes a cleaner global search space

• **Many improvements to search UI as part of CoG development:**
  ‣ Admin interface to customize a project specific search
  ‣ User interface to search for data
  ‣ Data cart to store search results and invoke data services
In 2016, the MSWT will focus on the following tasks (of decreasing importance):

• Support deployment of Publishing and Search Services across the federation
  ‣ Support to ESGF node administrators during installation and data publication
  ‣ Monitor consistency of search results across the federation

• Implement metadata validation against Controlled Vocabularies (CVs)
  ‣ Support project-specific CV profiles

• Develop tools and services to support atomic metadata updates
  ‣ Evolving QC flags, PIDs, DOIs for datasets and files
  ‣ Attach new services to already published datasets

• Support tagging of datasets for multiple projects
  ‣ Searching across MIPs and searching only a specific MIP (WIP/CMIP6)

• Package standalone authorization service to be deployed on Index Node to authorize publishing operations
2016 Future Roadmap

- **Continuos upgrade to newest versions of Solr: 5.3.1, 6.x**
  - Develop tool for seamless migration of Solr indexes

- **Support partitioning of search space across multiple Virtual Organizations**
  - CMIP, ACME, etc... may want to be searched separately

- **Implement other changes/improvements to the Search back-end and front-end (CoG) as they are requested/vetted/prioritized by the community**

- **Review and expose documentation for users and node administrators**
  - Most importantly: search RESTful API

- **Possibly: release alternative Python-based software for publication**

- **Research usage of SolrCloud**
  - Many advantages: automatic replication and failover, performance, scalability, automatic distributed indexing
  - Problem: architected for internal nodes, not to replicate across remote nodes
• Concrete, usable implementation of CVs to validate data/metadata before they are published
  ‣ Must have an implementation format ASAP, cannot wait till the data are published
  ‣ Must coordinate ES-DOC and WIP separate efforts

• Federation-level policies for sharing/distributing metadata across indexes
  ‣ Publish data that does not need to be federated to the local shard
  ‣ Must control who can publish to high profile projects such as CMIP6, Obs4MIPs, ...
  ‣ Must comply with project requirements about data content, metadata completeness, directory structure, supporting tech notes, etc.

• Would welcome additional team members especially to take responsibility for metadata standards and implementation

• Need help with setting up the infrastructure for monitoring the consistency of distributed searches