ESGF F2F 2017 SUMMARY

Findings

- General opinion is that it's much more important to focus on a few things that must work flawlessly for CMIP6, instead of adding new functionality
  - Publishing, search, download, replication
- There is a need for an Integration Working Group with the following tasks:
  - Establish and enforce release timelines
  - Execute thorough testing of release candidates
- We do not need a node manager
  - Every application that exchanges information across peers is a potential risk and should be peer-reviewed by many people, and based on well known tested libraries
  - There are already open source solutions for distributed registries

Priorities

Short Term

Priorities for CMIP6

- Test and freeze a stable Data Node release that can be installed at all Tier-2 sites
- Must Establish a fixed release schedule
- Must establish validation procedure for ESGF releases
    - Matrix of institution X functionality checks
- Improve how we handle publishing or replicas across the federation
    - Perhaps have a central registry ?
- Evaluate PrePare performance and possibly enhance not to slow down publication
- Remove access control for downloading data for specific projects
    - Which projects ? CMIP6, CMIP5, obs4MIPS, CORDEX…
    - Maintain access control and registration for compute
    - Changes in the stack:
        - Change wget script generation for index node - simply URLs
            - Decide based on some configuration file ? Or use existing config files but deploy them on index node
        - Change access policies at all data nodes
- Harden and test replication
    - Define replication tests and automate them
        - Selection files for each Tier1 site (partially complete)
        - Automation of download tests via Synda
    - Tune network between data centers
        - Guided by test replication performance data
    - Document procedure for setting up DTN
        - Renew network architecture discussion - network architecture diagrams from all Tier1 sites
    - Publish data with DTN URLs
- Need to train data providers on how to publish CMIP6 into ESGF
- Establish a policy and procedure for removing a non-compliant data node from the federation
    - In cease they don't upgrade to a security fixing release within X days
    - In case they publish non-conformant data
- Establish **"Data and Service" challenges** to publish test CMIP6 data
    - Use production Federation across at least 3 Tier 1 nodes
    - Protect data from download via special download group
    - Use test CMIP6 data, multiply volume by 1000
    - Will stress system end-to-end including:
    - Scalability of publisher
    - Necessary changes to search facets

    - Replication
    - Will test integration with ES-DOC and other CMIP6 services
- Make sure we can handle CMIP6 data from Asia
    - Which data nodes and index node will be used ?
    - How do we replicate the data ?
- Enhance Lukasz monitoring tables with notification to node administrators
- Make sure that UI for CMIP6 can significantly feature selection by MIP era and scenario MIP
- Make sure that all CoG sites allow access to the same data - must review consistency
    - CMIP6, CMIP5
    - Endorsed CMIP6 MIPS
    - obs4MIPs, input4MIPs

Priorities for non-CMIP6 specific items

- Establish user group to review UI usability and make constructive suggestions
    - Must be implementable in the short term
    - Better if composed by external people, not ESGF developers
- Establish task force to review possible data download methods
    - Must evaluate versus easy to use FTP
    - Make sure that Globus downloads work at each site - see NCAR RDA example
- Cleanup the global search space
    - Remove test project and other garbage

Medium Term (Sometime in 2018)

- Include CWT compute engine in ESGF releases
- Possibly, republish all CMIP5 data
    - Would solve user complaints about not being able to download data by variable
    - Good stress test in preparation for CMIP6
- Migrate Globus infrastructure to GCS v5 (within 2018)
    - No more globus-url-copy
    - Distributed as RPM or Docker container
    - Must upgrade all ESGF IdPs to OpenID-Connect
- Evaluate whether we still want to support LAS
- Establish a system for better user support
    - Options: Stack Overflow, AskBot, JIRA
    - Seed site with FAQs from CoG
    - Must handle both technical and scientific questions
    - Must allocate funded personnel
    - Also improve error messages to reduce the number of user requests
- Improve documentation at all levels

    - Decide on centralized site - perhaps one for end users and one for developers
    - Youtube for younger generations - Social media strategy?
- Harvest information from search logs to understand the most common use cases

Longer Term

- Evaluate whether community is better served by having centralized project specific User Interfaces instead of a distributed portals based on geographic node
    - Possibly hosted on Cloud with agnostic domain such as esgf.org
    - Must plan for redundancy and automatic failover
    - Must establish cost model
- Start experimenting with Google Cloud Platform
    - Discuss cost model
    - Perhaps setup a full node, or perhaps a Search Super Node only
    - Coordinate with CWT as hybrid cloud computing to be integrated
- Improve consistency of metadata by requiring that each project specifies a schema before data can be published
    - Build a publisher that will use the schema to validate data upon publishing
- Evaluate whether we want to replace CoG with search-only web interface like impact-4-climate portal
- Extend and document API access to the data to support future Virtual Labs and in Europe the Open Science Cloud concept
- Prepare plan for Machine Learning tools
    - Review state of the art
    - Prepare and distribute sandboxes

Release Process

- The Data Challenges are run on the Test Federation.
- The Test Federation is protected so that it is not exposed to users or to potential operational security issues

- A step to make a release go to production is that the security scan is performed. This means there may be several Data Challenges on the Test federation before a release is pushed to production. Dan says that its a two week period to scan for security.
- Tests may also be checked on the production if appropriate.
- Appropriate Documentation must be available with the full release.
- Communication to the user community must be included as part of the full release so that the Exec/nodes(?) are clear on the message. We need to be clear on who/how communication will take place - assemble a contact list.
- We need a roll-back procedure for production.

Major items on the Release schedule

Core services

- Publishing
- Searching
- Pids
- Downloading
    - HTTP
    - Globus
- Replication
    -

Other timeline:

- Regridding and subsetting by end of the year

Integrated Test and Data Challenges

Once per month to have an integrated test.

Data Challenge number	Work packages Freeze Date and	ESGF Data Challenge date	Target Production Release at Tier 1

	package	commences	nodes
DC1		15 January	15 February ?
DC2			
DC3			
CMIP6 production ready			*1 June 2018*

Use a dedicated ESGF Slack Channel for out-of-band coordination on the tests.
Must execute security scan before deploying in production

Data Challenge	Tests included	Notes
DC1	T1.1,T1.2,	
DC2		X10 data size
DC3		X100 data size

The first tests can be run by the competent developers of the components. But the requirement is to write down the tests and what success means. The second (or later set of tests) are to be run independently of the developers. Where end-users are the target audience then exemplar end-users must be sought (at least from the WIP) to ensure the tests cover the requirement.

Definition of the Data Challenges tests

Test	**Description**	**Measure of Success**	**Specific issues to test**
T1	Publication of CMIP6 test data	Verify full publication of all files into DB; TDS; Solr	
T1.1	PrePARE Scanning of CMIP6 data	PrePARE is performant for appropriately sized data	PrePARE memory leak needs checking.
T1.2	Publication step	Publication is performant for appropriately sized data	

T1.3	PID assignment DOI assignment	PID resolution possible	Ensure test flag set at all sites
T1.4	ES-DOC Simulation scanning (happens within esg-publisher)	JSON records have been received by ES-DOC server.	Identified that conda environment needs fixing; possible global attrs issue
T1.5	Errata service for updated data		
T1.6	Revoke data at a site	Unpublish from the federation.	
T1.7	Publish new version	Publish an updated collection of datasets (with new version date)	
T2	Synda Replication based on CMIP6 across Test Federation		
T2.1	Replication at scale across the main Tier 1 nodes		
T2.2	Replication through the replicated node		
T2.3	Per variable (“dataset”) replication and republication		
T2.4			
T3	Uniformity of Search from a User perspective		Canada as friendly user starting July? For every user perspective?
T3.1	CoG search ease of use	Karl to provide some details?	
T3.2	Ensure the results are not confusing from a user perspective.	Shows correct results if data version has been updated at the main node, but not yet updated at other nodes	
T3.2	Ensure consistent results of a search across the nodes		
T3.3			
T4	Download of data from a User perspective		Karl to provide?

T4.1	Wget works		
T4.1.1	Works with current authentication		This requires also testing of the dashboard stats to make sure works for the stats that need to be reported.
T4.1.2	Works without authenticated access on data		
T4.2	Globus		
T4.3	Validate download logging	Entries are added to access_logging table	
T4.4	Data aggregation works		
T4.5	Data subsetting works		
T5	Dashboard validation		
T5.1	Properly running (health check from CMCC)		
T5.2	Properly provide federation statistics		

Site Leads:

- IPSL: Guillaume
- CEDA: Ag
- DKRZ: Stephan
- GFDL: Serguey
- NCI: Jon
- JAMSTEC: ?
- LLNL: Sasha

Service/software Leads:

- PID: Tobias
- Errata: Atef
- ES-DOC “cdf2cim” integration: Ag (working with David Hassell, Mark Greenslade)
- Search, CoG: Luca

- Publisher: Sasha
- Synda: Guillaume
- esgprep: Guillaume
- PrePARE: Denis (PCMDI)
- Installer: William H., Sasha
- Dashboard:
    -