Second ENES Data Task Force Meeting
June 8th and 9th, 2016 in Paris at IPSL

Version June 23rd, 2016: Michael L. with contributions from Sebastien, Frank, Stephan, Martin an d Antonio

Day 1: Coordination of CMIP6/ESGF related activities (June 8th, start at 13:30 h)

Attendees : Francesca Guglielmo (IPSL), Guillaume Levavasseur (IPSL), Mark Greenslade (IPSL), Wim Som de Cerf (KNMI), Antonio S. Cofiño (UoC), Stéphane Sénési (CNRM), Grigory Nikulin (SMHI), Philip Kershaw (CEDA), Ingo Bethke (UNI), Sébastien Denvil (IPSL), Martin Juckes (CEDA), Atef Ben Nasser (IPSL), Pierre Antoine Bretonnière (BSC), Michael Kolax (SMHI), Stephan Kindermann (DKRZ), Michael Lautenschlager (DKRZ), Frank Toussaint (DKRZ).

Status new ESGF Services: Citation, PID, Annotation / Errata, CIM, Replication (Michael L., presentation attached)

Session started with an introduction of the CMIP6/AR6 data life cycle and its three phases: data production, project or community phase and the bibliometric phase. The ESGF CMIP6 data publication will be conditioned by checkpoints for the new services:

- PID syntax is correct
- Citation information is complete
- Errata information available when publishing a new version
- Further_info_url is correct

The implementation of the new ESGF services seems to be within schedule. Details had been presented in individual presentations. Focus of this meeting was the initial ESGF publication of CMIP6 data.

Demo and status of the errata service (Guilliaume and Atef)

Only modellers and data provider can create and modify issues in the errata service. Each replaced version should carry an entry in the errata service. Issues are pushed on GitHub in private mode. GitHub offers us the authentication/authorisation system.
It is planned that ES-DOC consumes and indexes issues and provides tools and services to consult errata issues. Errata issues will be discoverable by /institute/model/experiment/MIP. Given a list of tracking_ids a summary of the issues affecting data files will be received and an information whether a new version is available.

Action: The errata service deployed in beta testing this fall. This includes access to the service via ESGF GOC. PID service and ES-DOC.

Demo and status of SYNDA for the ESGF replication service (Sebastien)

Sébastien highlighted a discussion on esgf-devel scoping ESGF client requirements
https://docs.google.com/document/d/1ecgqPIqMshaGja476SQ5m4JfcCaiTNJlrAd0NACaicY/edit#heading=h.7s4q1zbs0qnr
Synda allows to discover what is available in ESGF from the command line and allows to specify synchronous and asynchronous data replication. Search features include an inference module on facet parameter from facet value.

- synda search 20160101-20161231 "Air Temperature" -f

- synda get tasmax_day_FGOALS-s2_piControl_r1i1p1_20160101-20161231.nc # synchronous
- synda install tasmax_day_FGOALS-s2_piControl_r1i1p1_20160101-20161231.nc # asynchronous

Synda can use http or gridftp in dependence of ESGF data nodes. Post-processing modules can be triggered upon datasets download completion (QC, or ESGF replica publication ...). The current version is synda 3.4.1.

Action: Synda is ready for CMIP6 replication as soon as the CMIP6 DRS will be fixed. Synda will be included into the ESGF data node implementation SW stack for CMIP6.

Status of PID (Persistent Identifier) and QA (Quality Assurance) service (Stephan, presentation attached)

The prototype of the handle service in place. CMOR3 is prepared to create tracking-ids in PID syntax with UUID. The PID client will be included in the ESGF data node implementation SW stack and will push CMIP6 PIDs to the global handle service as part of the ESGF publisher. A RabbitMQ in front of the PID registration process is installed to cache heavy PID loads. A second exchange node in Europe to addition to DKRZ and another in the US would improve the reliability.
The CMIP5 QA at DKRZ has been completely redesigned. A separate CF checker module has been included together with components for checking basic netCDF file publication requirements. A light weight QA checking the minimum requirements for CMIP6 will be provided by CEDA on basis of the CMIP5 version.

Action: For the PID service 5 months of beta testing is planned. The service then will start this fall. The DKRZ QA will be completed by end of 2016 together with a web front-end for spot checks. Beta testing will start in summer together with the finalisation of the CMIP6 global attributes and the related controlled vocabularies (CV).

Action: Determine hosting of a second ENES PID exchange node.

Status of ES-DOC/CIM (Mark)

CIM2 is currently in its final phase of design. Tools are designed to adapt to CIM2 representation. Activities around the Scientific Properties CVs are important and on the critical path. CMIP6 experiment descriptions for ES-DOC has been defined by Charlotte Pascoe and are under review. Model descriptions for CMIP6 will start from already existing CIM documents for CMIP5. A utility to auto-document simulations from NetCDF files is going to be prototyped. The CMIP6 data citation information will be linked at the level the further_info_url landing page. The errata service will be linked into ES-DOC but the granularity is different from that of the further_info_url. Documentation on CMIP6 forcing should be integrated in ESDOC as well.
Action: After 5 months of beta testing starting the service is planned this fall.

CMIP6 data request (Martin)

Specification of variables, output requirements, experiment specification suffers from the CMIP6 complexity of 22 endorsed MIPS. The Data request issued by WGCM to modelling centres to have a more coherent archive in CMIP6. The status of the process can be obtained from the URL https://earthsystemcog.org/projects/wip/CMIP6DataRequest (or, equivalently, w3id.org/cmip6dr). Three different types have to be mapped, CMOR variables, MIP variables and requested variables. Not all variables have descriptions but 95 % have.
Because of the complexity of the request and the large number of contributors it is hard to set a specific date for completeness. There will be version controlled updates. It is now clear that getting the variable definitions finalised is a priority, while details of mapping data requirements to specific experiments and objectives is less urgent.

Action: European modelling group meeting for feedback might help. Martin will prepare list of questions and material related to the topic of missing variable definitions that might be passed to some modelling groups. A follow-up telco with the ENES DTF is suggested.

Climate4Impact presentation has postponed to Thursday and the discussion of requirements for ESGF publication of CMIP6 files has already been integrated into the earlier discussion of QA tools.

Day 2: ENES DTF (June 9th, end 12 h)

Attendees: Sylvie Joussaume (IPSL), Francesca Guglielmo (IPSL), Guillaume Levavasseur (IPSL), Mark Greenslade (IPSL), Wim Som de Cerf (KNMI), Antonio S. Cofiño (UoC), Nikolai Kadygrov (IPSL), Christian Pagé (CERFACS), Grigory Nikulin (SMHI), Philip Kershaw (CEDA), Ingo Bethke (UNI), Sébastien Denvil (IPSL), Martin Juckes (CEDA), Pierre Antoine Bretonnière (BSC), Michael Kolax (SMHI), Stephan Kindermann (DKRZ), Michael Lautenschlager (DKRZ), Frank Toussaint (DKRZ).

Demo and status of Climat4Impact portal (Wim and Christian)

A nice user interface had been presented. Login is possible with ESGf accounts and DKRZ and IPSL should be added as identity provider for the portal. For CMIP6 the errata information will be made available via ES-DOC for requested data entities. The downscaling portal form the University of Cantabria is included.
The increasing data complexity for CMIP6 requires forward thinking to prioritize search results. CORDEX has access statistics for the most popular variables (50% of data requests for 10 variables).
Action: CMIP5 access statistics should be evaluated to infer prioritization of CMIP6 search results.

Review action items from April 29th, 2015

The action item to prepare coherent presentations on existing infrastructure embedded nationally and present them on a small workshop end of 2015 is still open. It is important for a clearer view on the ENES data infrastructure and its existence after the end of the ISENE-2 project.
Action: The coherent infrastructure presentation will be revisited.
Michael L. agreed to chair the ENES DTF.

European coordination CMIP6 data management

Netcdf4

Netcdf4 compressed is the requested CMIP6 format but it is not precisely written down in WIP documents. This is important for data volume estimates in the ESGF federation. The volume has strong implications on the CMIP6 replication strategy. A volume estimate is planned by the WIP for end of June. NetCDF files with unlimited dimension have performance issues on client-side tools processing those files. The impact in NetCDF4 is lower than in NetCDF3 files.
Action: Antonio will send reference to the group with respect to best practices on chunking NetCDF4 files.

CDNOT

CDNOT has started operation. Monthly telcos are scheduled. Security contacts at the data node sites have been identified. PGP keys has been exchanged to enable secured channel of communication. Next step is to get everybody up to speed with respect to CMIP6/WIP documents & requirements

CMIP6 tier1 and tier 2 specification for European data nodes

The ESGF XC send a draft specification of tier 1 and tier 2CMIP6 data nodes to the WIP and CDNOT. Resource requirements assume that CMIP6 core data is 10 times more than for CMIP5 and a tier 1 data node should be able to store a complete copy as for CMIP5. This result in a disk space requirement of about 20 PB and related assumptions for network bandwidth for replication and compute resource for data processing. At least for European ESGF data nodes the disk space requirement seems to be unrealistic which has implications for a CMIP6 replication strategy. Requirements for bandwidth and compute resources need more rationales and explanations as well as the burden from CMIP6 on tier 2 sites. Specification of tier 1 and tier 2 should include level of service, HW (spinning disk, compute, network, tapes), support tier2, and support for providers.

Action: Specification of CMIP6 tier 1 and tier 2 is expected form CDNOT and from the ENES DTF.

Action: Balaji will send CMIP6 data volume estimates which are crucial for resource specifications tier 1 and tier2 as well as for the data replication strategy.

CMIP6 replication strategy

Action: Michael et al. to start from the CMIP6 data volume to propose a replication strategy

Action: Michael & Sébastien. ENES strategy wrt replication be included in Tiers1/2 requirements

Action: European partners to put on the table what they could contribute

Sustainability of ENES data infrastructure

In kind contributions

Not discussed

RM ODP (https://en.wikipedia.org/wiki/RM-ODP) contribution from institutes (refer to minutes from last DTF general discussion item 4)

A parallel view to RM OPD is provided by ISO. The ISO framework has three levels of detail under each heading. Martin presented a CEDA description using ISO while Stephan prepared an RM ODP for DKRZ separating component and deployment diagrams.

There was general agreement to produce a simple version of institutional descriptions within ENES DTF to highlight complementary within the ENES data infrastructure.

Action: Each institute will create a description of their data infrastructure and use that as a basis for discussion about how we might define an ENES data infrastructure.

Action: Martin provides information on the ISO framework to describe ENES data infrastructure sites.

IS-ENES2 Deliverable 11.1: ENES Data Activity Delivery plan

D11.1 is actually complete, but the DTF should review it to see if we are sticking to the plan.

Not discussed.

Support within European Open Science Cloud

Not discussed

ENES and EUDAT

Not discussed

ENES data infrastructure as part of ESGF and alternatives

Sylvie pointed out the probable funding gap until 2018 after the end of IS-ENES2 and that ENES is not yet not on the ESFRI roadmap, and so does not qualify for consideration in projects such as the EOSC.

She gave an update on ENES infrastructure strategy workshop which is planned in the week of 24th October in UK. The ENES data infrastructure should be a topic in the workshop which sets time constraints to action items in previous topics in this meeting.

Michael closed the meeting at 12:15. The work in the ENES DTF will be continued with more regular telephone conferences and by using the already existing Redmine wiki (https://redmine.dkrz.de/projects/enes-data-task-force/wiki) as information platform.