Minutes and Actions, CORDEX, Hamburg 30th/31st May, 2012

P2P

-- index nodes: 1st at DKRZ, later in 2012 at DMI and BADC; 2013 LIU.
Index nodes lead data management policy development and enforcement.
-- data nodes: Running at DKRZ, IPSL, soon at BADC and DMI; 2013 LIU, Cape Town.
-- authorisation server -- straightforward: Ole can send terms to Hans.
-- Configuration issues: DKRZ has configuration file with all necessary information.
-- Consider using harvested CMIP5 catalogue if distributed search has problems;

Roadmap

-- Stable URL for 1st CORDEX user interface with search for all CORDEX data.
enes_dn1.dkrz.de
-- First final data from SMHI early June.
-- start publicising the link, mid june; CORDEX data only; With warnings on user interface stability;
-- implement version control with drslib -- June, possibly after initial publication.
-- When PCMDI P2P index node is stable: update our index nodes to put CORDEX and CMIP5 data in same search interface;
-- Data expected: IPSL end of year; MOHC late summer; Croatia ready at DMI; MPI end of year; DMI soon; Uni. Cantab -- some format issues; SMHI some data ready, more imminent;

Policy

-- Talk to NCAR -- Michael to start discussion with Steve Worley;
-- Standardised usage logs -- using data node db;
-- Memorandum of understanding for data node managers and team manager;
  -- Enforce specified compliance before publication;
  -- All data to be published with checksums;
  -- No data changes without new version;
  -- Use specified directory structure;
  -- Inform someone if node manager changes;
  -- Respond to user queries (nominate person to respond);
  -- Prompt action to correct publication errors;
-- Enforcement: checking.
  -- Can check for checksums and directory structure in catalogues of new datasets (easy);
  -- Checking all catalogues for illicit changes (difficult);
  -- Checking all data (impossible) -- rely on user reports;
-- Enforcement:
  -- remove a node from our search interfaces;
  -- notify WGCM;
  -- publish warnings for users;
-- Mailing list
  -- setup list for CORDEX data node managers (after signing MoU). cordex-dn@jiscmail.ac.uk
-- Help desk
  -- As CMIP5: helpdesk at BADC, list of experts (categories security, wget scripts, science);

-- Errata page is desirable and room for improvement -- e.g. data services of IS-ENES2. Maybe use off-the-shelf system (e.g. OSQA) or COG (new system designed by Sylvia and Luca for community discussions). User forum vs. FAQ.

Available capacity
DMI: 15TB now, more later
DKRZ: 1-200TB;
BADC: 1-200TB;
IPSL: 100TB;

Croatian data -- 1 simulation (Era-interim, Europe, 20yr, “44”, compressed) = 41GB.
Africa, “44”, uncompressed, Era-interim, 30yr: 100GB;
Historical simulation, 55yr, Africa: 178GB;
Scenario, 95yr, uncompressed; 356GB;

Africa: 10 era-interim + 15 * (hist + 2*rcp): 4000 model years, 3GB/model year ⇒ 12TB (uncompressed).
Euro-44: 1GB/model year (uncompressed).
Euro-11: 14GB/model year (uncompressed). 10 groups.
All for core and tier-1.
Tier-2 not requested as yet. 4 to 4.5 times (core+tier1) → 63GB/model year, 77GB total (Euro-11).

Ole and Grigory to make more precise estimate of Euro-44 expectations.

Data ingest and quality control

-- requirements: data specification document on cordex.dmi.dk (26th Jan, 2012);
-- Boundaries of interpolated data regions need to be agreed and specified precisely;
-- enforcement of specified time segments: not in generic SMHI script, nor in qc_basic.py; Can be added to qc_basic.py
-- vocabularies: would be useful to prepare more structured lists [for scripts] for those not yet covered; e.g. driving GCM [derived from <institute>-<model> for CMIP5 models] -- additional names to be registered with Ole (also for RCM names), experiment names need to be inserted in MIP tables: syntax with version number -- some groups will use rxiypz.
experiments: historical, evaluation, rcp45, rcp85, rcp26, rcp60
-- Publish at each variable or frequency level; variable level desirable to give more detailed information. Implications for publication and search efficiency? DKRZ will look into efficiency and appearance issues;
-- Version control in directory structure. Omit “latest”. Use “files” and “vyyyymmdd” and links (at frequency of variable level). Use hard or soft as preferred by local managers.

day/files/tas_20120101/file1.nc
/v20120101/tas/<link1>

or

tas/files/v20120101/file1.nc

/v20120101/<link1>
Documentation.
-- CIM: try to get questionnaire ready next year; cleaning up access to CMIP5 CIM is a priority. Meeting in Autumn (with NCAR) dedicated to setting some documentation roadmap; review potential of developments from PIMMS (more light weight, more flexible);
QC developments / roadmap
-- Basic compliance check before publication. Errors must be fixed before publication.
-- Data published to "cordex" authorisation group (cf "cmip5_research"); (terms of use as agreed -- to be posted on cordex.dmi.dk);
-- Errors reported by users or archive groups → errata page/database : no short term solution in existing infrastructure. For community site, would need moderation to avoid unacceptable comments appearing on an official site; Initial step is to collect information through helpdesk, post relevant comments on DMI errata page. Link to index notes to be discussed later.
-- When new version published -- policy on retention is: aim to keep data for medium term (3 years). Long term issue is curation → may only return one version. Long term at DKRZ generally curation subject to detailed documentation and DataCite DOI publication.
Ingest coordination
-- Transition of data from modelling centres.
-- Ole and Grigory to collate existing information from groups -- and place on ENES portal;
-- Contact Ole, register point of contact, model name and institute name, intentions (# simulations, timescale), agreement to terms of use and level of service;
-- Assign groups to data nodes (DKRZ, IPSL, LIU, DMI, STFC) -- plan tbd;
To national institute where applicable, SMHI → DKRZ initially; Other to DMI for initial compliance checks -- possibly send data to DKRZ if DMI node is not ready;
-- Modeling group run compliance checker and transfer data and MD5 checksums (disk or ftp or gridftp); [need to check access routes to each institute, and bandwidth];
-- Receiving institute verify checksums;
-- SMHI can offer support for use of the SMHI script; PCMDI does provide support for CMOR -- IPSL support for CORDEX CMOR configuration issues;
-- Data node runs compliance checker and publishes data;
-- Replication when stable;
Curation and DOIs
-- Initial phase is data sharing; groups should keep backup until long term storage terms are agreed.

ACTIONS
* Doodle for telco in late June [Martin];
* Verify SMHI script, compliance checker and publication sequence [Karin, Grigory, Estani];
* Need to add agreement on sharing email to terms of use; [Martin, Ole]
* Set up protected CORDEX workspace on ENES portal; [Stephan]
* Test P2P for possible problems with variable level publication [DKRZ];
* Clarify "terms of use" (is there a research only category?) [Ole];
* Set up authorisation server on an index node (enes_dn1.dkrz.de, and "P2P peer group" [DKRZ];
* Publish DKRZ P2P configuration files;
* SMHI script posted on DMI CORDEX site with guidance page [Grigory, Ole];

* STFC compliance checker posted on DMI CORDEX site with guidance page [Martin, Ole];
* Ask P2P developers to add option for facet guidance to interface;
* Draft MoU between data node managers and team leaders [Michael and Martin];
* Transfer of publication ready data from SMHI to DKRZ [Grigory];
* Publish SMHI data at DKRZ, with appropriate warning about interface stability [DKRZ];
* Re-configure drslib for cordex specification (including versioning) --- by end of June [Stephen];
* Talk to NCAR -- Michael to start discussion with Steve Worley;
* Cron job to harvest information from usage logs and file system [Michael, Sebastien -- to ask Sandro to tackle this -- June 4th telco];
* Set up help-desk (cordex-helpdesk@stfc.ac.uk) at BADC (with errata status in workflow), with list of experts [BADC];
* Draw up list of science experts [Ole];
* Errata page at cordex.dmi.dk [Ole];
* Add link to CORDEX help deskto index node (replace “contact ESGF”) [DKRZ];
* Collate existing information from groups (including contact for user data queries) -- and place on cordex.dmi.uk; [Ole, Grigory]
* Create page explaining ingestion procedure to modelling groups [Ole];
* Vocabularies spreadsheet [Martin];

Late 2012

* Put CORDEX and CMIP5 into same search interface, with links to both help desks [DKRZ];
* Script to check catalogues for checksums, and for changes at fixed version [IPSL];
* Create page with information on projected archive contributions [Ole, Grigory];

2013

* Integrate CERA database into index node to create entry point for DOIs [DKRZ];