SMHI

Data Reference Syntax – Governing Standards within Climate Research Data archived in the ESGF

Michael Kolax
Swedish Meteorological and Hydrological Institute

Motivation for a DRS within CMIP5

- In CMIP5 the decision was made to archive data within a grid federation of data nodes hosted at contributing centers.
- Within this federated archive data should be unambiguously identified.
- Catalog and file based infrastructure as for instance THREDDS and OPeNDAP where chosen for the ESGF
- In the federation data sets are replicated. That should not cause a change in file names or catalog structures , exception is the URL part identifying the data node hosting the replica.
- Users should find predictable names for core data sets to ease the data access.
- Download scripts should be easy to adapt for changes in one attribute as for instance host or variable names.

CMIP5 Data reference Syntax (DRS) and Controlled Vocabularies
Karl E. Taylor, V. Balaji, Steve Hankin, Martin Juckes, Bryan Lawrance, and Stephen Pascoe
http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

What is the DRS ?

- The Data Reference Syntax is a naming system to be used within files, directories, metadata and URL's to identify data sets wherever they might be located within the distributed ESGF archive.
- DRS is heavily dependent on controlled vocabularies and CF standard names.

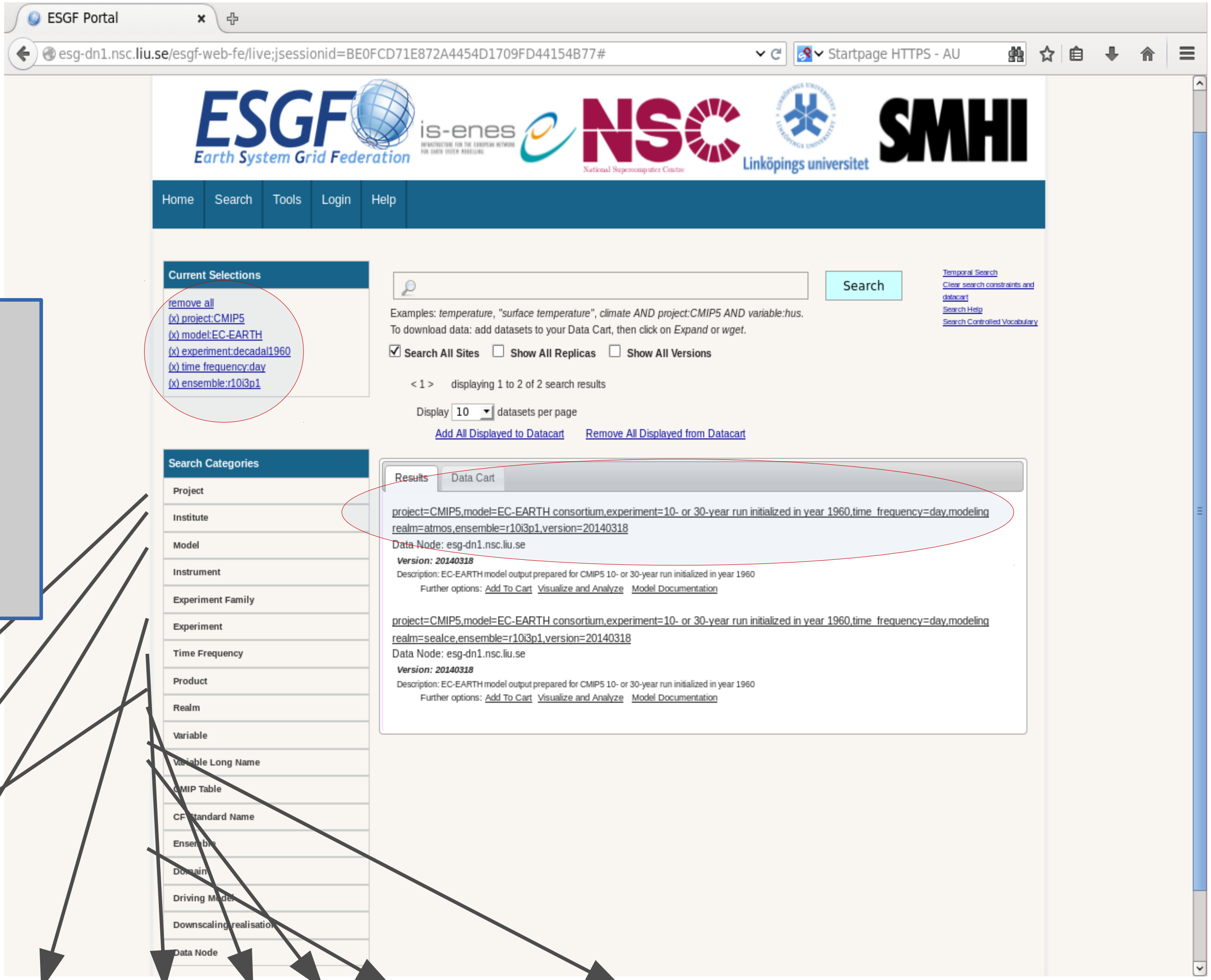
  (Controlled vocabulary schemes mandate the use of predefined, authorized terms that have been pre-selected by the designer of the vocabulary.)
- DRS aims to be as flexible as possible while at the same time giving clear and structured conventions for data management.
- DRS is checked by various different quality or compliance checking tools and recent projects demand the checking before data is published.
- The controlled vocabulary building up the DRS is showing up in search interfaces and therefore should be brief an clear.
- DRS usually does not make any rules about more general aspects of data handling as for instance checksumming or the use of tracking ID's.

Example:

cmip5/output/smhi/EC-EARTH/decadal1960/day/atmos/day/r10i3p1/v20140318/clt/clt_day_EC-EARTH_decadal1960_r10i3p1_19610101-19701231.nc

ESGF Search Interface and DRS

DRS elements are exposed in the ESGF search interface.



ESGF Portal
esg-dn1.nsc.liu.se/esgf-web-fe/live;jsessionid=BE0FCD71E872A4454D1709FD44154B77#
Startpage HTTPS - AU
ESGF Earth System Grid Federation
is-enes
NSC National Supercomputer Centre
Linköpings universitet
SMHI
Home
Search
Tools
Login
Help

Current Selections

remove all
(x) project:CMIP5
(x) model:EC-EARTH
(x) experiment:decadal1960
(x) time frequency:day
(x) ensemble:r10i3p1

Search Categories

- Project
- Institute
- Model
- Instrument
- Experiment Family
- Experiment
- Time Frequency
- Product
- Realm
- Variable
- Variable Long Name
- CMIP Table
- CF Standard Name
- Ensemble
- Domain
- Driving Model
- Downscaling realisation
- Data Node

Search

Temporal Search
Clear search constraints and datacart
Search Help
Search Controlled Vocabulary

Examples: *temperature*, "*surface temperature*", *climate AND project:CMIP5 AND variable:hus*.
To download data: add datasets to your Data Cart, then click on *Expand* or *wget*.

☑ Search All Sites ☐ Show All Replicas ☐ Show All Versions

< 1 > displaying 1 to 2 of 2 search results

Display 10 datasets per page

Add All Displayed to Datacart Remove All Displayed from Datacart

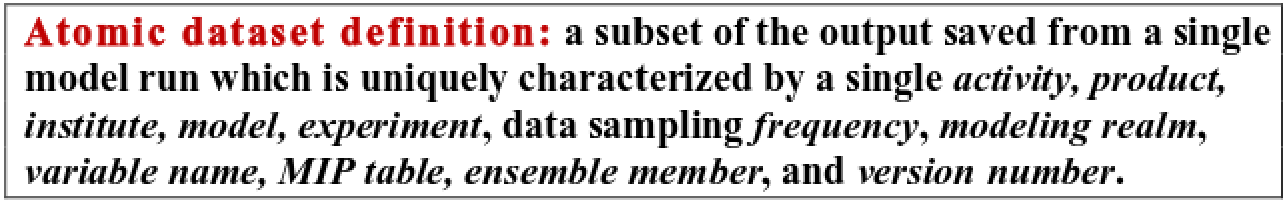
Results | Data Cart

project=CMIP5,model=EC-EARTH consortium,experiment=10- or 30-year run initialized in year 1960,time_frequency=day,modeling realm=atmos,ensemble=r10i3p1,version=20140318
Data Node: esg-dn1.nsc.liu.se
Version: 20140318
Description: EC-EARTH model output prepared for CMIP5 10- or 30-year run initialized in year 1960
Further options: Add To Cart Visualize and Analyze Model Documentation

project=CMIP5,model=EC-EARTH consortium,experiment=10- or 30-year run initialized in year 1960,time_frequency=day,modeling realm=seaIce,ensemble=r10i3p1,version=20140318
Data Node: esg-dn1.nsc.liu.se
Version: 20140318
Description: EC-EARTH model output prepared for CMIP5 10- or 30-year run initialized in year 1960
Further options: Add To Cart Visualize and Analyze Model Documentation

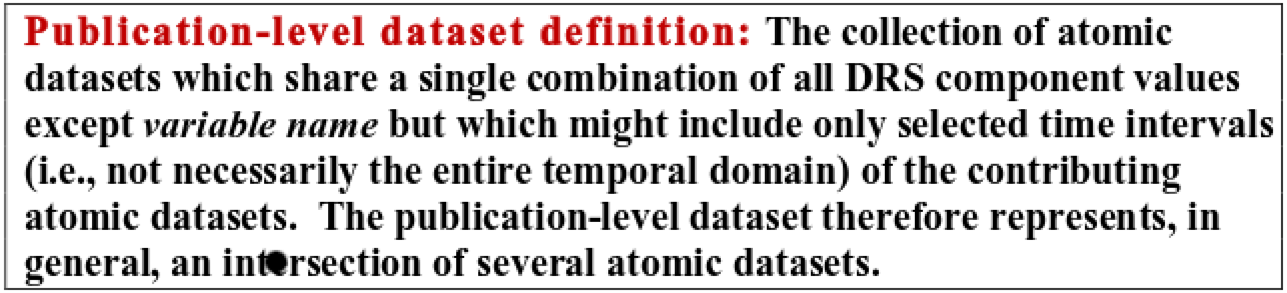
cmip5/output/smhi/EC-EARTH/decadal1960/day/atmos/day/r10i3p1/v20140318/clt/clt_day_EC-EARTH_decadal1960_r10i3p1_19610101-19701231.nc

DRS Founding Definitions CMIP5



Atomic dataset definition: a subset of the output saved from a single model run which is uniquely characterized by a single *activity, product, institute, model, experiment*, data sampling *frequency*, *modeling realm, variable name, MIP table, ensemble member*, and *version number*.

An atomic data set contains the entire spatio-temporal domain for one variable.
Besides the last two DRS members all vocabulary is controlled.



Publication-level dataset definition: The collection of atomic datasets which share a single combination of all DRS component values except *variable name* but which might include only selected time intervals (i.e., not necessarily the entire temporal domain) of the contributing atomic datasets. The publication-level dataset therefore represents, in general, an int●rsection of several atomic datasets.

The CMIP5 DRS

- CMIP directory structure (as created by CMOR) :

  <activity>/<product>/<institute>/<model>/<experiment>/<frequency>/<modeling realm>/<variable name>/<ensemble member>/

  ESGF data node directory structure:

  <activity>/<product>/<institute>/<model>/<experiment>/<frequency>/<modeling realm>/<MIP table>/<ensemble member>/<version number>/<variable name>/

- File name:

  <variable name>_<MIP table>_<model>_<experiment>_<ensemble member>_<temporal subset>_<geographical info>.nc

Controlled Vocabulary

- by experiment leader
- by data provider approved and controlled by experiment leader

Controlled Structure

- chosen by data provider

Activity: Model intercomparison och data collection activity (eg CMIP5)
Product: Characterizes a type of data (eg output , derived)
Institute: Name of the Institute responsible for the Model run (eg : UKMO) , in case of CMIP5 authorized by PCMDI
This institute name may differ to official institute_id (global netcdf attribute) which may contain characters not permitted in DRS (eg : blank, period)
Model: Model identifier (eg : EC-EARTH), in case of CMIP5 approved by PCMDI
Experiment: Identifies the experiment (eg: rcp45) , as CV - Short Name of Experiment
Frequency: Indicator of interval between individual time-samples (mon, fx)
Modeling realm : Indicates which high level modeling component is of interest for this dataset (eg: atmos. Ocean)
Variable Name: The Variable name has to be taken from the MIP table (CF standard)
MIP table: Describes the MIP table of the dataset. For CMIP MIP Tables are collecting fields sampled only at single frequency
Ensemble member: A formatted string describing aspects of closely related simulations (aka 'rip' number).
Version Number: The version number is the character v followed by an integer, in CMIP representing the publication date.

DRS elements and their counterpart in global NetCDF Attributes

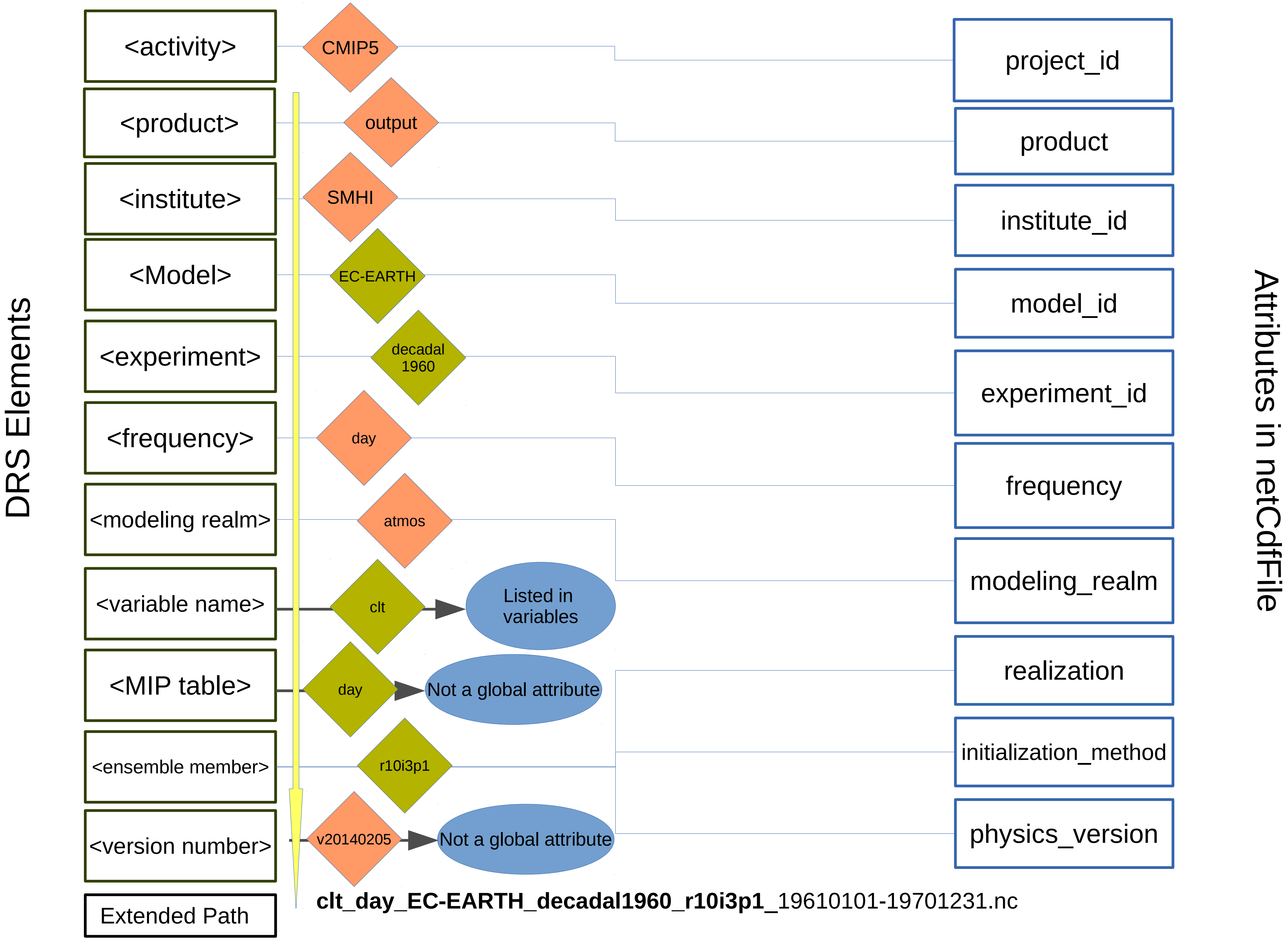
- Most DRS Elements have an associated global Attribute in NetCDF files. This is in most cases a one to one relation. Exception from that is for instance the DRS element <ensemble member> in CMIP5.
- DRS puts hard constraints on what characters might be used to build up the fields. To avoid misinterpretation eg : in URLs, only the following characters are allowed for building up component values :

  a-z A-Z 0-9 -
- The official CMIP5 institute id might include other characters which might lead to the effect that the institute_id as NetCDF attribute contains characters not allowed in the DRS element <institute>

(CMIP5 DRS, anything ESG-published like that ?)

CMIP5 DRS to Global Attribute Relation

file name construction



DRS Elements

<activity>

<product>

<institute>

<Model>

<experiment>

<frequency>

<modeling realm>

<variable name>

<MIP table>

<ensemble member>

<version number>

Extended Path

CMIP5

output

SMHI

EC-EARTH

decadal 1960

day

atmos

clt

Listed in variables

day

Not a global attribute

r10i3p1

v20140205

Not a global attribute

project_id

product

institute_id

model_id

experiment_id

frequency

modeling_realm

realization

initialization_method

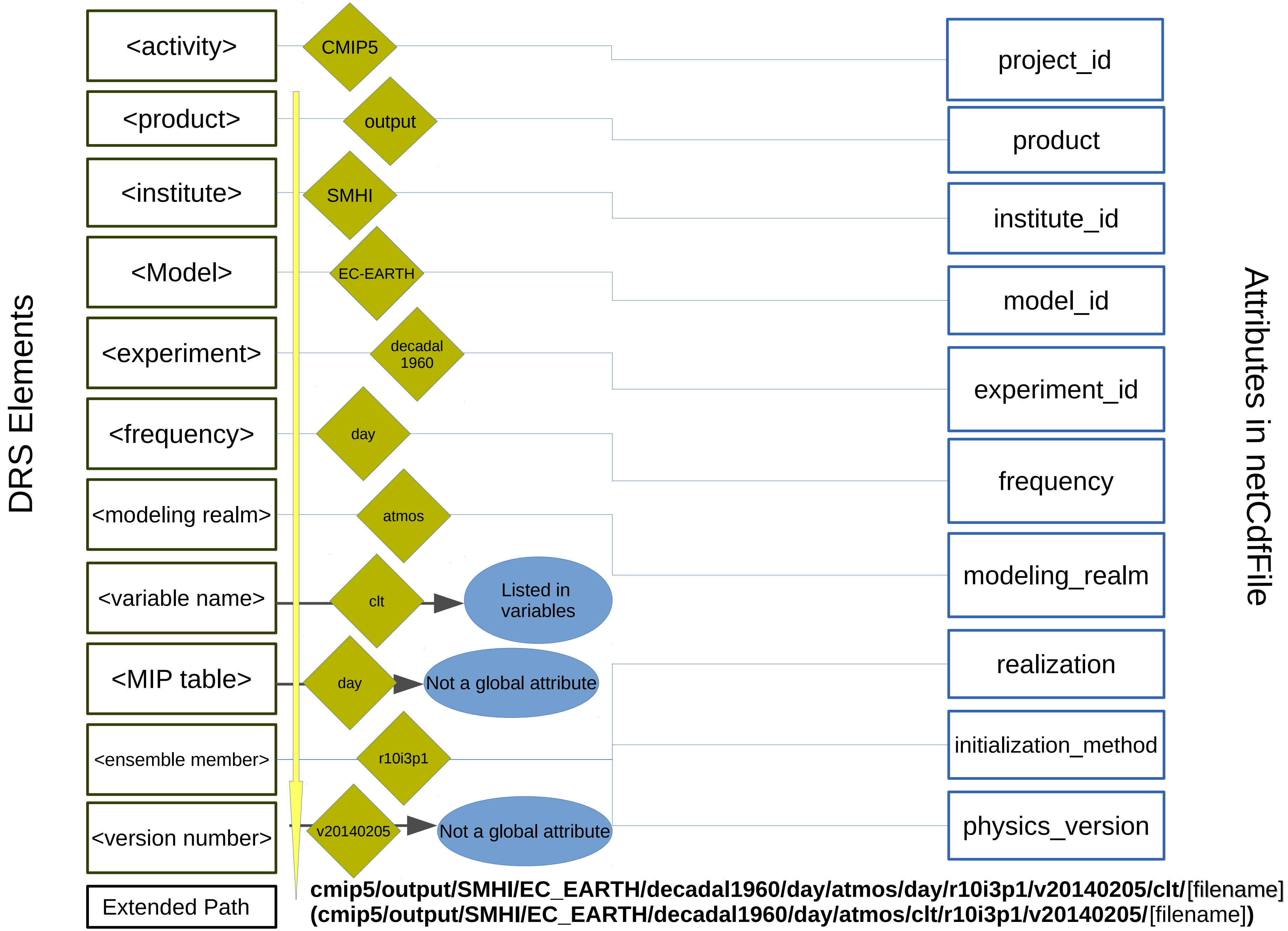
physics_version

Attributes in netCdfFile

clt_day_EC-EARTH_decadal1960_r10i3p1_19610101-19701231.nc

CMIP5 DRS to Global Attribute Relation

directory construction



cmip5/output/SMHI/EC_EARTH/decadal1960/day/atmos/day/r10i3p1/v20140205/clt/[filename]
(cmip5/output/SMHI/EC_EARTH/decadal1960/day/atmos/clt/r10i3p1/v20140205/[filename]**)**

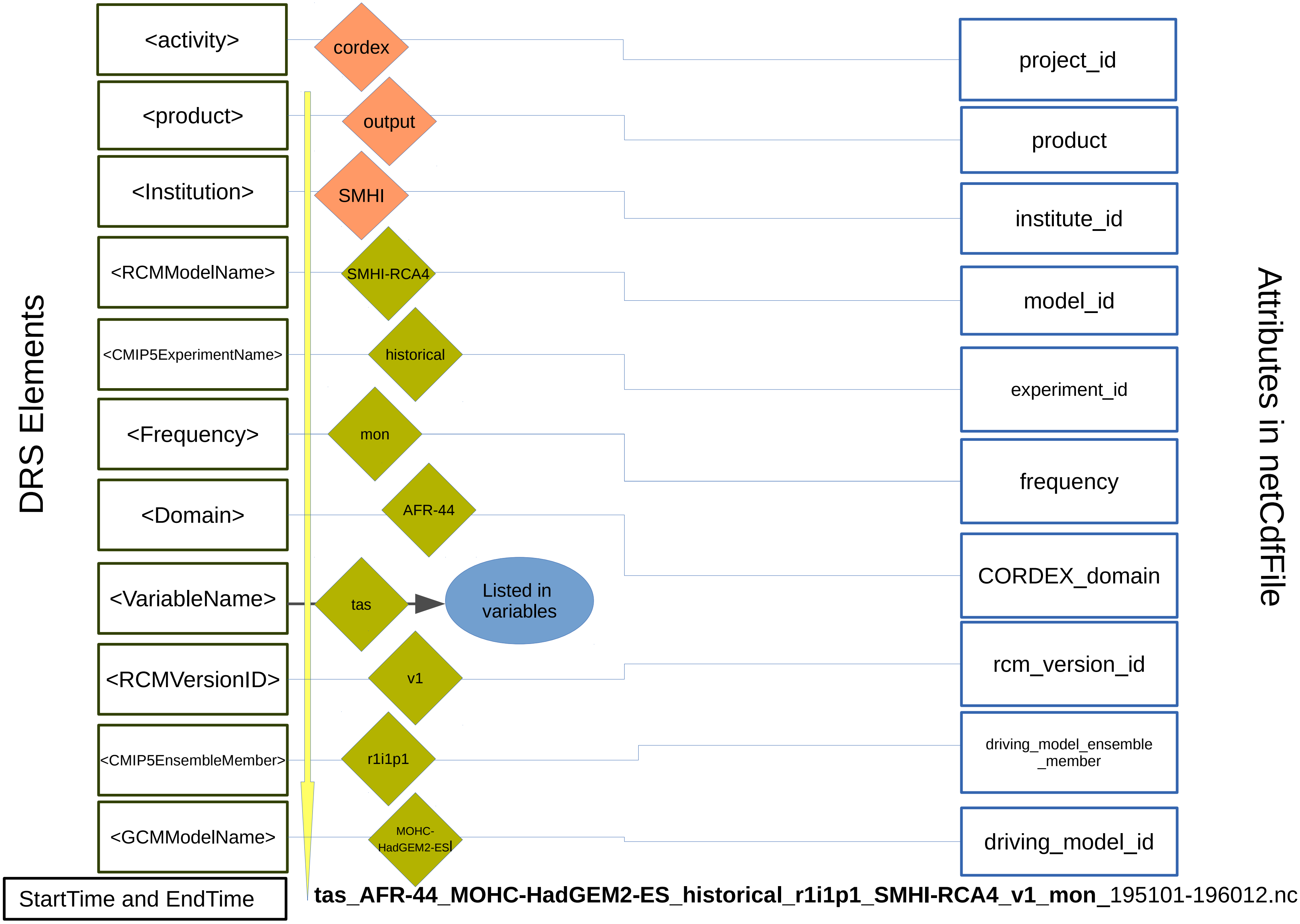
The CORDEX DRS

CORDEX as regional climate downscaling experiment had needs to cover different information within the DRS than global modeling activities.

- CORDEX DRS deploys CMIP5 as default where own definitions are missing
- Domain was included
- RCMModelName was included
- GCMModelName was included instead of Model
- A RCMVersionID was included
- CMIP5EnsembleMember was included instead of Ensemble Member
- CMIP5ExperimentName was included instead of ExperimentName
- Modeling Realm was removed
- MIP Table disappeared

CORDEX DRS to Global Attribute Relation

file name construction



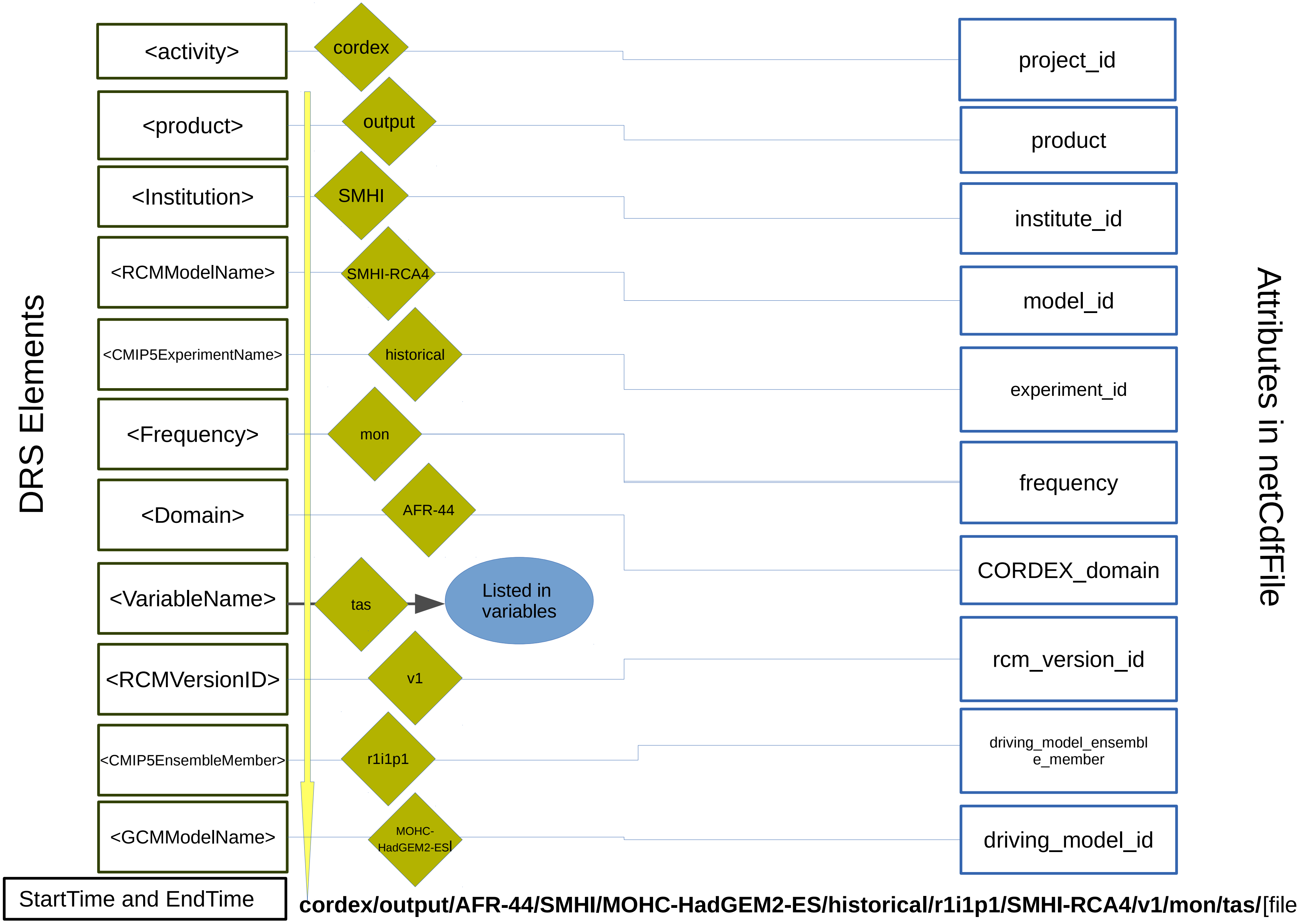
DRS Elements		Attributes in netCdfFile
<activity>	cordex	project_id
<product>	output	product
<Institution>	SMHI	institute_id
<RCMModelName>	SMHI-RCA4	model_id
<CMIP5ExperimentName>	historical	experiment_id
<Frequency>	mon	frequency
<Domain>	AFR-44	CORDEX_domain
<VariableName>	tas	Listed in variables
<RCMVersionID>	v1	rcm_version_id
<CMIP5EnsembleMember>	r1i1p1	driving_model_ensemble_member
<GCMModelName>	MOHC-HadGEM2-ES	driving_model_id

StartTime and EndTime

tas_AFR-44_MOHC-HadGEM2-ES_historical_r1i1p1_SMHI-RCA4_v1_mon_195101-196012.nc

CORDEX DRS to Global Attribute Relation

directory construction



DRS Elements		Attributes in netCdfFile
<activity>	cordex	project_id
<product>	output	product
<Institution>	SMHI	institute_id
<RCMModelName>	SMHI-RCA4	model_id
<CMIP5ExperimentName>	historical	experiment_id
<Frequency>	mon	frequency
<Domain>	AFR-44	CORDEX_domain
<VariableName>	tas	Listed in variables
<RCMVersionID>	v1	rcm_version_id
<CMIP5EnsembleMember>	r1i1p1	driving_model_ensembl e_member
<GCMModelName>	MOHC-HadGEM2-ESl	driving_model_id
StartTime and EndTime		

cordex/output/AFR-44/SMHI/MOHC-HadGEM2-ES/historical/r1i1p1/SMHI-RCA4/v1/mon/tas/[filename]

DRS for Regional Reanalysis Data

- CLIP-C is currently working on making EURO4M data available through the ESGF interface. A proposed DRS for regional Reanalysis is aiming to cover even ensembles of regional reanalysis as planned in UERRA.

Many DRS fields resemble similar entries in CORDEX DRS:

```
VariableName_Domain_RADrivingName_Experiment_RADrivingEnsembleMember_RAModelName_RAEnsembleMember_Frequency[_StartTime---EndTime].nc
```

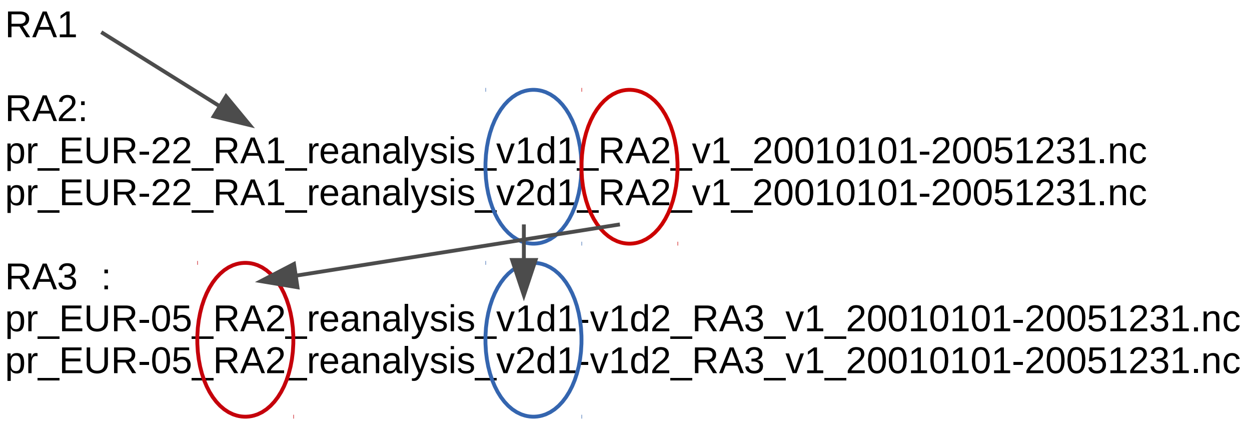
```
<activity>/<product>/<Domain>/<Institute>/<RADrivingName>/<Experiment>/<RADrivingEnsembleMember>/
<RAModelName>/<RAEnsembleMember>/<Frequency>/<VariableName>
```

One of the challenges is to build a in theory unlimited nested structure into the limited number of fields long DRS ...

Proposed DRS for Regional Reanalysis Data

- The Details of the sequence of driving RA's shall be provided only within the NetCdf attribute “driving_ensemble_description”.
- The nesting level and ensemble member information is included in in DRS element RADrivingEnsembleMember

RADrivingEnsemblemember identifies the ensemble member and sequence the driving RA in form of v<N>d<M> (V<N>ensemble member , d<M> sequence number of driving RA).



RA1

RA2:
pr_EUR-22_RA1_reanalysis_v1d1_RA2_v1_20010101-20051231.nc
pr_EUR-22_RA1_reanalysis_v2d1_RA2_v1_20010101-20051231.nc

RA3 :
pr_EUR-05_RA2_reanalysis_v1d1-v1d2_RA3_v1_20010101-20051231.nc
pr_EUR-05_RA2_reanalysis_v2d1-v1d2_RA3_v1_20010101-20051231.nc

Example: ECMWF-ERAINT->SMHI-HIRLAM->SMHI-MESAN
pr_EUR-22_ECMWF-ERAINT_reanalysis_v1d1_SMHI-HIRLAM_v1_20010101-20051231.nc
pr_EUR-05_SMHI-HIRLAM_reanalysis_v1d1-v1d2_SMHI-MESAN_v1_20010101-20051231.nc

Some General Remarks

- Project specific DRS development has proven to be an essential working procedure to meet the needs of a federated file based archive as the ESGF.
- So far most DRS fields can be kept human readable.
- We have successfully published data having more complex hierarchy than CMIP5 data and succeeded to provide a DRS (CORDEX).
- To establish a DRS can be a tedious effort, especially when actions are taken to late in a project.
- Well build DRS and correct meta data becomes even more important as we build services on top ESGF.