EUCLIPSE Deliverable 1.6

Reprocessed version of EUCLIPSE model data products
for long-term archiving within WDCC
beyond the runtime of the project

H. Ramthun
German Climate Computing Centre, DKRZ GmbH, Hamburg, Germany

June 2014

Index

1. Introduction

EUCLIPSE (**E**uropean **U**nion **Cl**oud **I**ntercomparison, **P**rocess **S**tudy & **E**valuation) is a project funded by the European Union under theme 9 "Environment" of Framework Program 7 of the European Union. It is designed to improve the evaluation, understanding and description of the role of clouds in the Earth's climate with a focus on the cloud feedback in a warming climate.

In parallel to the *EUCLIPSE* project all participating modelling centres also produced model data output for the model intercomparison project CMIP5 (Coupled Model Intercomparison Project Phase 5). A core part of these data are the base for the Fifth Assessment Report (AR5) of the *IPCC* (Intergovernmental Panel on Climate Change). Most data for the Euclipse project have been processed together with that data production.

During the project phase 50 TByte of physical storage is available at the DKRZ for temporal data storage, project data exchange and additional data processing. For data publication and distribution the *ESGF* (Earth System Grid Federation) gateway was installed to make data available inside the *ESGF*.

Finally the Euclipse project output data will be moved into the Long Term Archive (LTA) at the WDCC (World Data Center for Climate) at the DKRZ. Further information about the LTA workflow is found in the LTA handbook available on the DKRZ webpage (http://www.dkrz.de/Nutzerportal-en/doku/hpss/lta-doku).

2. Euclipse project output

A lot of the Euclipse project output data has been generated together with the model runs for the CMIP5 of the *IPCC*.

In Euclipse work package 1 (WP1) the **COSP** (**C**FMIP **O**bservational **S**imulator **P**ackage) software was used to generate output. The software has been released for implementation in the EUCLIPSE climate models as version 1.2.2 by Alejandro Bodas-Salcedo. More information about *CFMIP* (Cloud Feedback Model Intercomparison Project) and the software can be found on the *CFMIP* (www.cfmip.net/).

Two additional experiment series have produced Euclipse output data: COOKIE and SPOOKIE.

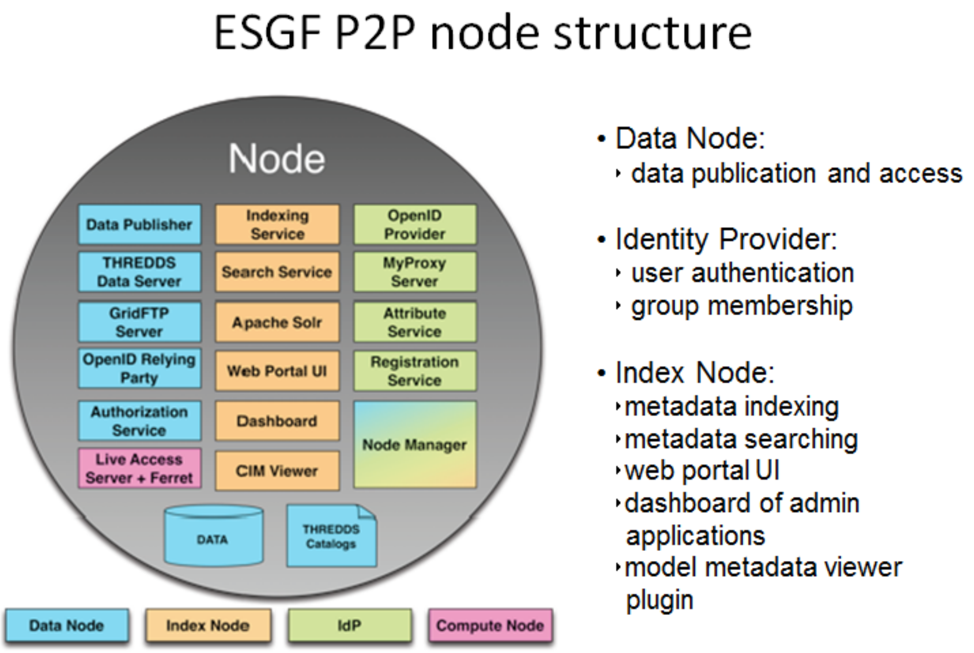
COOKIE, the **C**louds **O**n-**O**ff **K**limate **I**ntercomparison **E**xperiment, is designed to expand on the AMIP (Atmospheric Model Intercomparison Project) and aqua-planet subset of the CFMIP simulations conducted as part of CMIP5 and mainly represents the Euclipse WP4. The main characteristic of the COOKIE experiments is that the impact of the clouds on the radiation is switched-off. A detailed publication about COOKIE can be found on the Euclipse web page here: *COOKIE*.

SPOOKIE, the **S**elected **P**rocess **O**n/**O**ff **K**limate **I**ntercomparison **E**xperiment is very similar to the COOKIE experiment. In the SPOOKIE experiments the convection has been switched off.

The delivery of COOKIE/SPOOKIE data will still be ongoing when this deliverable 1.6 has been finished.

3. ESGF (Earth System Grid Federation)

For the CMIP5 report the immense amount of more than 2 PByte data has been produced. To deal with such large data volume with its hundreds of variables in millions of files a new distributed software structure was developed and implemented during the project time. The development process is still ongoing and will be continuously extended with new functions like server side processing to reduce the amount of transported data. Several people from all modelling and data centres joined this federation to contribute to its success. The main outcome of the *ESGF* is the 'peer to peer' (P2P) gateway. The idea behind this structure is to make it possible to install at any site the appropriate functionality to serve the local data requirements.



ESGF P2P node structure

- Data Node:
  - data publication and access
- Identity Provider:
  - user authentication
  - group membership
- Index Node:
  - metadata indexing
  - metadata searching
  - web portal UI
  - dashboard of admin applications
  - model metadata viewer plugin

Figure 1 ESGF node components

Each node consists of one or more services as shown in Figure 1. The type of node can change and is only distinct by the installed services. Nodes are grouped into so-called peer groups. Information about the available data is exchanged between all nodes in a peer group. The general interaction structure between two p2p gateways is shown in Figure 2.

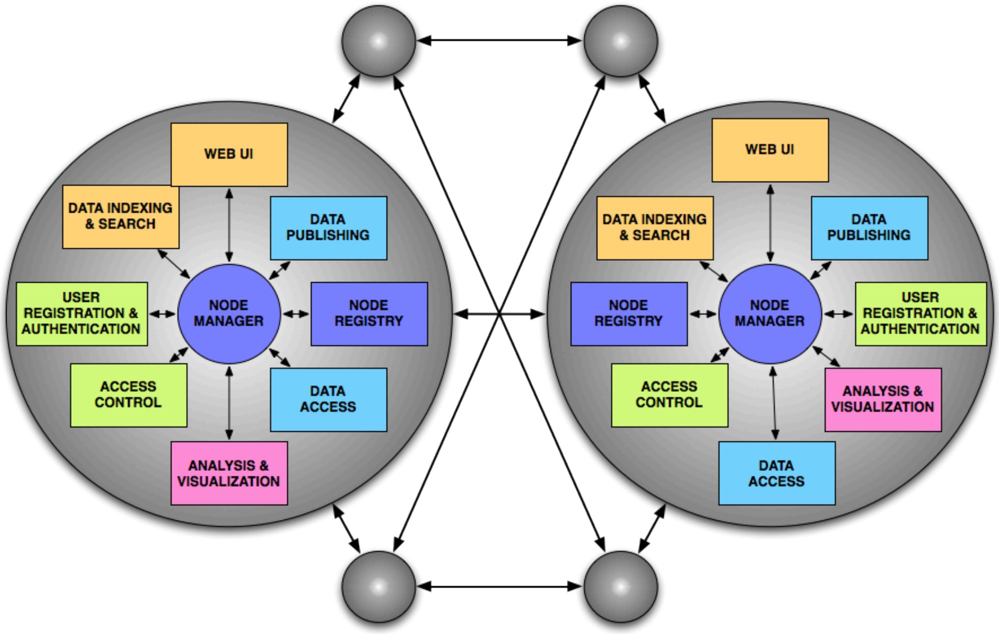


Figure 2 ESGF Gateway Interactions

All modelling centres, which participate in the CMIP5 data production, have published their data to one of the P2P (peer to peer) index nodes (gateway) of the ESGF federation. They share the same production peer group.

A detailed documentation about the complete ESGF is found in the *ESGF* wiki.

An ESGF node is a system of installed components. Each administrator can choose during installation a type of node: Identity Provider (IDP), Index Node (index), Data Node (data), Compute Node (compute). The node type is changeable by adding or removing components in a later update process. Each node consists of a group of low level services which are deployed during the installation of the process.

All data are published to a data node. Each data node is connected to an index node to make the data available inside the system. The index nodes are grouped by peer groups and data of one particular data node is only cycled inside a peer group.

Figure 3 shows the plan of the actually available nodes (data) and gateways (index) in the ESGF production federation and their connection to data centres and models.

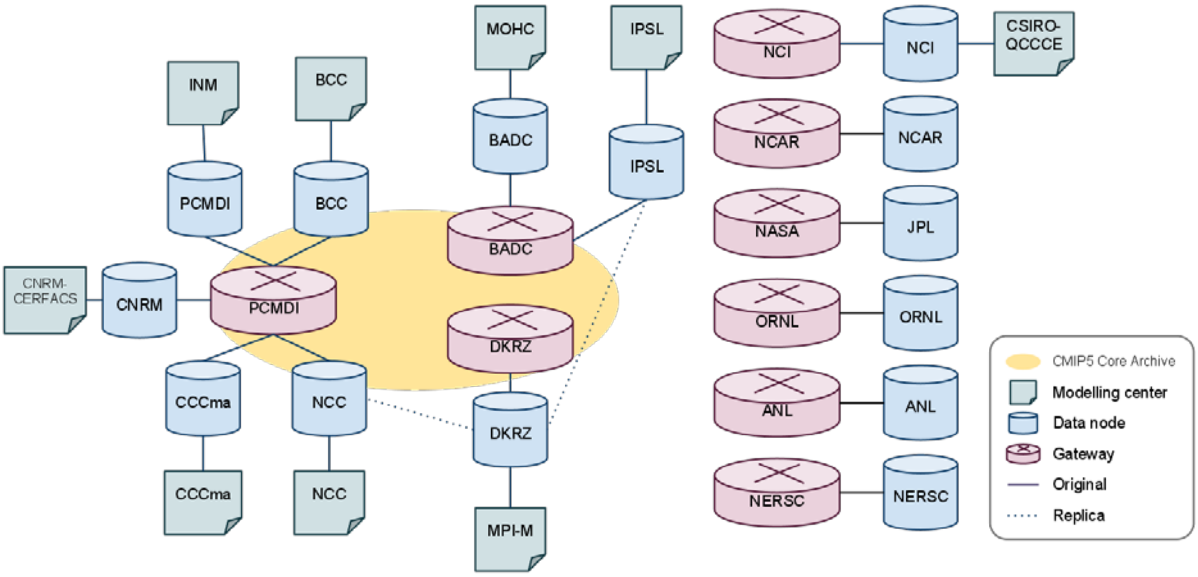


Figure 3 Overview of ESGF Nodes

4. Euclipse Data at DKRZ

The DKRZ is a German facility to provide a high performance computing platform, high sophisticated data management and other services for the climate science community.

Detailed documentation about DKRZ can be found on the *DKRZ Webpage*.

The role of the DKRZ as a Euclipse project partner is to be responsible for a platform which facilitates storing data temporarily, that enables an exchange of data between institutes or to publish data to make them available for a wider community. After the expiration date of the project, Euclipse data will be moved to the DKRZ long term archive.

The Euclipse server is integrated in the DKRZ infrastructure as shown in Figure 4.

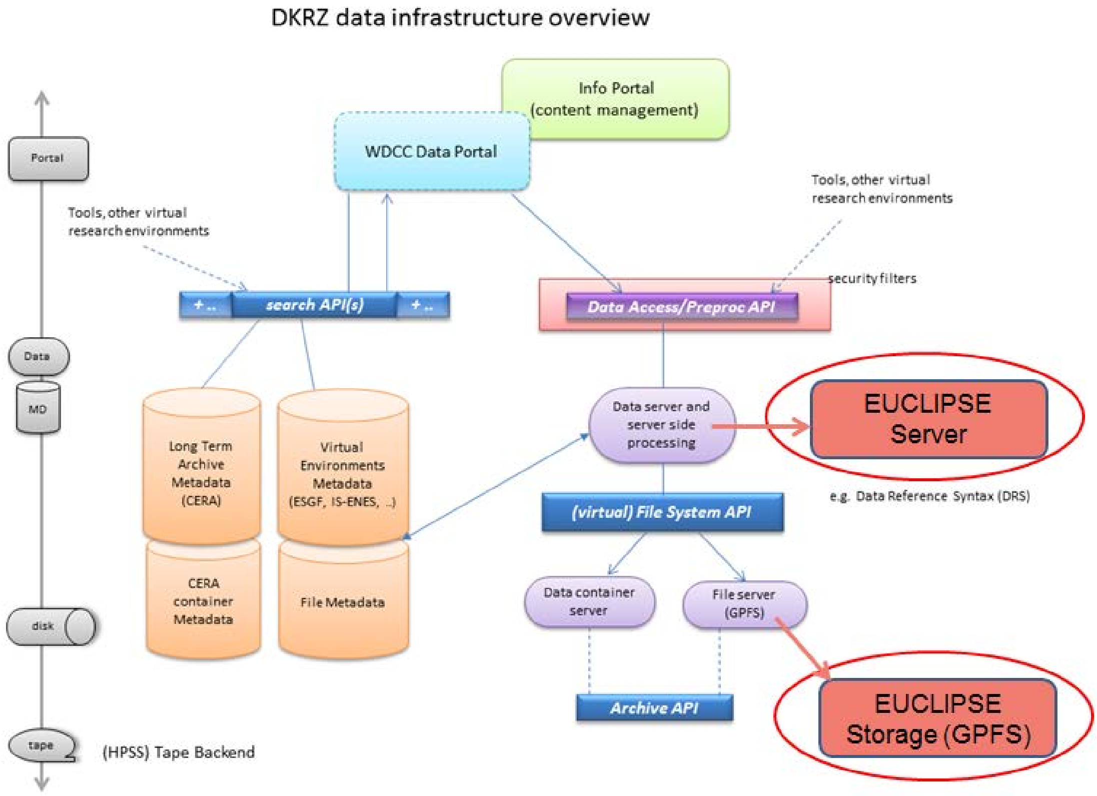


Figure 4 Euclipse facilities inside DKRZ

The server is a separated DELL system with 8 cores and 31.5 GByte main memory.

4.1 Euclipse storage

A 50 TByte partition of disc storage is reserved for the usage by all Euclipse project partners to exchange data and make them available for partners in other projects. Accessing the Euclipse storage area by sftp requires a standard DKRZ account. All these DKRZ accounts are connected to the Euclipse project where all interested people are open to register as a member. With this registration they are able to access all available project facilities.

All CMIP5 model output data of MPI-M models are located in the same file system as all replicated CMIP5 data from other institutes. So any processing capability e.g. using monthly data of amip* experiments of other models is facilitated.

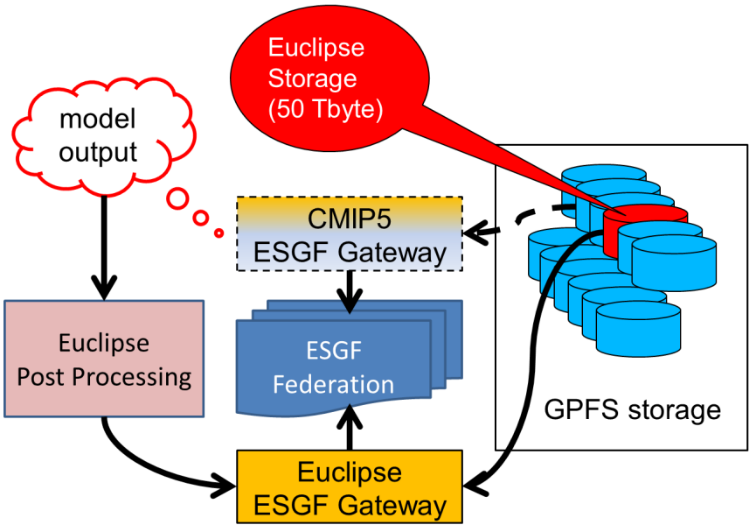


Figure 5 File System Storage – ESGF

4.2 Euclipse processing at WDCC

The Euclipse project aims at an improvement of understanding the role of clouds in climate modelling.

4.2.1 CFMIP processing using COSP

One implementation workflow is the usage of the COSP simulator (refer to WP1 of the Euclipse project). The output data processed by COSP allows a comparison with observed (satellite) data and thus an evaluation.

Additional processing (see below 4.2.4) has been done with CMIP5 data at DKRZ and is being stored in the Euclipse storage part of the DKRZ file system which is the GPFS of IBM (*IBM General Parallel File System*).

4.2.2 COOKIE processing

One specific result of the Euclipse processing is a new experiment described at *COOKIE,* and which was developed by Euclipse WP4). This experiment results in some new extensions in CMIP5_Amon and CMIP5_day tables:

Short description	experiment
'offAMIP'	'offamip'
'4xCO2 offAMIP'	'offamip4xCO2'
'offAMIP plus 4K anomaly'	'offamip4K'
'offaqua planet control'	'offaquaControl'
'4xCO2 offaqua planet'	'offaqua4xCO2'
'offaqua planet plus 4K anomaly'	'offaqua4K'

4.2.3 SPOOKIE processing

Some institutes (MIROC, MRI, CNRM and MPI-M) provide yet one more block of output data. Several other institutes also plan to perform SPOOKIE experiments and provide their data. The SPOOKIE data are quite similar to the COOKIE data (see chapter 4.2.2), the main difference being that they would be called convoffamip, convoffaqua, etc.

4.2.4 cdo processing

One other implementation workflow offered to the researchers as an extended service is shown in Figure 6: processing of monthly data of CMIP5 amip* experiments to monthly mean data over a sequence of years with the cdo (*climate data operators*).

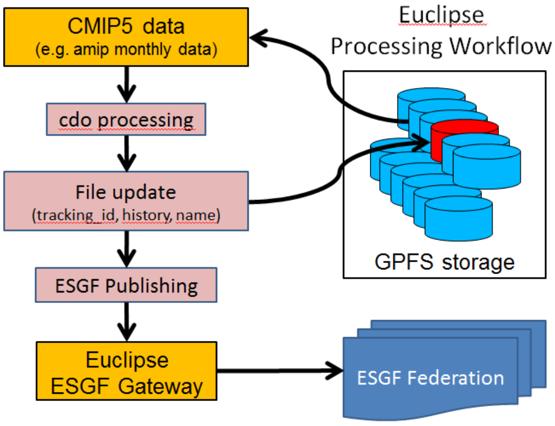


Figure 6 cdo Processing Workflow

This processing is simply done by the cdo operator 'yearmean'. This processing step results in an annual mean of monthly data. This output has only been produced for CMIP5 amip* experiments.

The agreement inside the CMIP5 community is to supply a new tracking_id to any new file. This newly generated tracking_id is written together with update history into the file header to document the processing steps. Also agreed inside CMIP5 are the DRS (**D**ata **R**eference **S**yntax) naming convention of files. The frequency part of the filename is changed from 'mon' to 'yr' in order to comply with the DRS.

This kind of processing is to be replaced in the future by the WPS (Web Processing Service) to perform e.g. cdo or other workflows.

4.3 Euclipse publishing

All Euclipse data, which have been produced by one of the processing steps, are published into the ESGF federation (see, for example, the blue rectangles in Figures 5 through 7). All COOKIE and SPOOKIE data have been generated with regard to the rules of CMIP5 data creation e.g. to meet the DRS. For the additional cdo processing this must also be ensured e.g. each file must have a new tracking_id. Data publishing into the ESGF federation as a new product has been done with the same ESGF publisher tool as used inside CMIP5 (please refer to chapter 3).

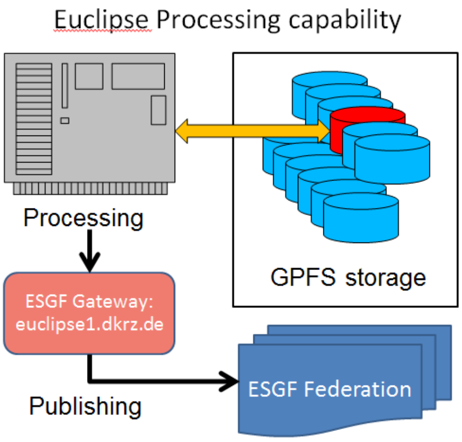


Figure 7 Publishing of processing products

Inside the ESGF there exist several peer groups of data production centres like, for example ‘production’ or ‘test’, indicating their purpose. The published Euclipse data is shown and accessible in the ‘production’ peer group.

4.4 Euclipse archiving

One main task is to move Euclipse project data into the LTA (*long term archive*) of the WDCC (*World Data Center for Climate*) with the CERA (*Climate and Environmental Retrieval and Archive*) database as storage for the metadata.

Therefore all steps required for DKRZ projects and described in the LTA handbook, must be also be performed: deciding which data should be archived, describing the data by metadata that are put into the CERA database and as a last step uploading the data into the database and move them to the HPSS tape library where they will be available for at least 10 years.

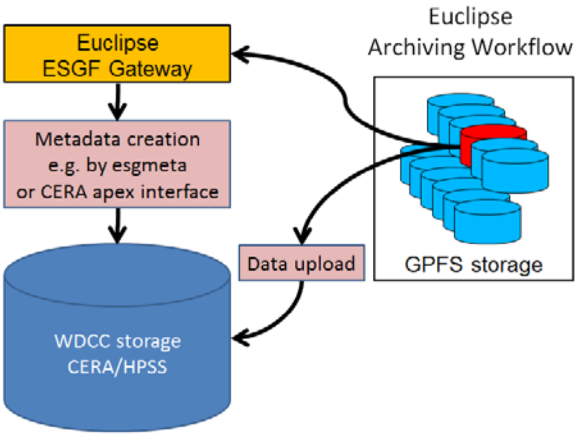


Figure 8 Archiving data

Creating metadata and uploading the data makes them available at the CERA portal of the WDCC (CERA (WDCC portal)).

All processed Euclipse output data have already been archived in the WDCC. To do this the following steps have been performed:

- Processing with new tracking_id as described before (this step is only done for the reprocessed amip* data with cdo processing)
- Publishing the datasets into the Euclipse1 ESGF index node
- Writing CERA XML metadata derived from the ESGF metadata
- Uploading these XML metadata into the CERA database (the WDCC data storage)
- Uploading the data itself to the tape archive

For more details about long term archiving of climate data at DKRZ please refer to the DKRZ LTA handbook: DKRZ LTA handbook.

5. Conclusions

Data storage of Euclipse model output data is available at DKRZ. Several data by the Euclipse project were produced together with the AR5/CMIP5 output. Additional data processing has been applied to CMIP5 data to provide an extra data collection beyond CMIP5. Processing is based on available tools like *climate data operators* or *netCDF operator* or other software. For example, the additional cfmip processing in Euclipse of CMIP5 data has been done by several modelling centres.

At DKRZ additional processing was carried out using common tools like cdo (09). For example the operator 'yearmean' was used.

All results were published on the Euclipse P2P gateway (http://euclipse.dkrz.de) into the ESGF federation inside the 'production' peer group. The 'publication' peer group then makes the data available to a broader community.

Euclipse project output data are archived in the WDCC *(12)* and will stay available for at least 10 years.

6. Acknowledgements

EUCLIPSE is funded by the EU 7th Framework as an e-infrastructure project (# 244067).

We are grateful to Mark Webb from the UK Met Office who had the idea for the additional cdo processing and Estanislao Gonzales who guided us to the secrets of the ESGF p2p gateway.

7. References

(01) Euclipse project page: EUCLIPSE, www.euclipse.eu
(02) CFMIP, www.cfmip.net/
(03) COOKIE and CREAM, www.euclipse.eu/wp4/wp4.html
(04) CMIP5 project page: CMIP5 , http://cmip-pcmdi.llnl.gov/cmip5/
(05) IPCC webpage: IPCC, http://www.ipcc.ch/
(06) ESGF wiki: ESGF, http://esgf.org/wiki/

(07) DKRZ webpage: DKRZ, www.dkrz.de

(08) IBM General Parallel File System, http://www-03.ibm.com/systems/software/gpfs/

(09) climate data operators, https://code.zmaw.de/projects/cdo

(10) netCDF operator, http://nco.sourceforge.net/

(11) long term archive, http://www.dkrz.de/daten-en/long_term_archiving

(12) World Data Center for Climate, http://www.dkrz.de/daten-en/wdcc

(13) Climate and Environmental Retrieval and Archive, http://www.dkrz.de/daten-en/cera

(14) COOKIE, http://www.euclipse.eu/downloads/Cookie.pdf

(15) CERA (WDCC portal) , http://cera-www.dkrz.de/CERA/