METAFOR Service Testing CIM Metadata in system METAFOR Deliverable 3.1 M33

PROJECT	
Project acronym	METAFOR
Project full title	Common Metadata for Climate Modelling Digital Repositories
Grant agreement no:	211753
Funding Scheme	Combination of Collaborative Projects & Coordination and Support Actions
Call Topic	INFRA-2007-1.2.1 Scientific Digital Repositories
DOCUMENT	
Deliverable	D3.1 Month 33
Deliverable Title	Service testing CIM metadata in system
Document Identifier	METAFOR-D3.1_M33
Date	December 20, 2010
Work Package	WP3 Service Content Creation & Testing
Authors	IPSL
Document Status	draft
Document Link	http://metaforclimate.eu/documents

Dissemination Level		
PU	Public	
PP	Restricted to other programmes participants	**X**
RE	Restricted to a group specified by the Consortium	
CO	Confidential	

Document History			
Version	Date	Comment	Author/Partner
0.1	December 20, 2010	First Draft	S. Denvil/IPSL
0.2	January 31, 2011	Second Draft	S. Denvil/IPSL
0.3	February 4, 2011	Third Draft	C. Pascoe/BADC H.Ramthum/DKRZ
0.4	February 12, 2011	Fourth Draft	M. Kolasinski/Climpact
1.0	February 15, 2011	Final Version	S. Denvil/IPSL

Abstract:

There is growing interest among researchers, policy-makers, businesses, and the general public in the potential impacts of climate change and to what extent undesirable consequences of climate change can be mitigated by reducing anthropogenic greenhouse gas (GHG) and specifically carbon emissions. This deliverable aims at demonstrating how

metafor services and tools can offer a framework to help interested parties addressing those critical issues.

Introduction

The Metafor Common Information Model, i.e. the CIM, is an ontology designed to become the ipso-facto standard for climate modeling related metadata. The CIM eco-system allows institutes to integrate the CIM into their day to day climate modeling processes. It achieves this by supporting various requirements: the ontology itself; validation; search; dissemination; integration (with other metadata platforms such as Earth System Grid). This deliverable aims at demonstrating how those CIM eco-system components can be orchestrated all together. The focus will thus be made on the integration aspect.

One of the prerequisites for all studies of climate change and its consequences is the existence of climate simulations at the global scale from which coherent atmospheric, oceanic or surface field time-series can be extracted for further modelling or statistical studies. Though such global simulations are produced for assessment by the IPCC, the constraints of such a wide international exercise do not allow the flexibility and interaction in the choice of scenarios and the storage of datasets featured in the ENSEMBLES project (CMIP5 will be different from what has been done before regarding this aspect). We choose the FP6 ENSEMBLES project as our main data source for this demonstration for four reasons:

- provide centennial and seasonal to decadal simulations
- provide high temporal resolution datasets covering the entire simulations ; well suited for impacts studies
- data are publicly available through standardized interfaces
- key WP3 partner were already involved in the ENSEMBLES project and thus have experience using those datasets

Metafor will provide services to allow users to populate a CIM repository with CIM instances complying with the CIM ontology. The repository may be populated manually or automatically using various tools provided by Metafor. For this demonstration the CIM repository will be populate using a derived CMIP5 questionnaire and the thredds2CIM tools with additional functionally.

The CMIP5 questionnaire is web a based interactive tool for creating CIM documents about the climate models that contribute to CMIP5, the CMIP5 Questionnaire also captures information about how those climate models were set up to run simulations for CMIP5 experiments. We will use a simplified version of the CMIP5 questionnaire to gather information about models used to perform seasonal to decadal and centennial simulation within the ENSEMBLES project.

As the CIM repository grows in size, it becomes important to allow users to efficiently and effectively perform semantically rich searches against the repository. Thus search tools and services will be made available to the community. Search complexity differs according to user type and ranges from simple to complex, thus Metafor must deliver a rich set of web-based search tools & services. The metafor search engine user interface will be integrated to the University Of Cantabria downscaling portal. This search engine will query METAFOR repository using METAFOR services architecture.

ENSEMBLES seasonal to decadal and centennial datasets

The ENSEMBLES project is funded by the European Commission (EC), and runs from September 2004 to December 2009. ENSEMBLES is a flagship project of the EC's 6th Framework Programme (EC FP), an integrated project under the thematic sub-priority 'Global Change and Ecosystems' (contract number GOCE-CT-2003-505539).

At the core of the ENSEMBLES integrated project was the development of the first global, high-resolution, ensemble based, modelling system for the prediction of climate change and its impacts. The Earth system models were combined into a multi-model ensemble system, with common output. For seasonal, decadal and centennial time-scales, RT2A purpose was to produce sets of climate simulations with several models and to provide the multimodel results needed for the other Research Themes. The results from RT2A were used for validation (RT5), studies of feedbacks in the Earth system (RT4), as well as boundary conditions and forcing fields for regional model simulations (RT3/RT2B). The simulations covered time-scales ranging from seasons to decades and centuries. Two streams of Global Climate Model (GCM) runs were produced: the first for the ensemble prediction system and the second using later models incorporating new features such as carbon cycle feedbacks. The development and running of the E1 stabilisation scenario was led by RT2A.

The ENSEMBLES project built ensemble prediction systems based on global climate models to generate projections of future climate on seasonal, decadal and multi-decadal time-scales. The scope included the assembly and testing of new global climate models, development and implementation of methods to represent the effects of uncertainties in the modelling of key physical, biological and chemical processes ('modelling uncertainties'), and the use of observations to initialise and constrain the projections. Seven European climate modelling centres ran GCMs under historic and four different scenario forcings (B1, A1B, A2, 1%CO2). All centres ran several realisations to create multi-simulation ensembles of most scenarios, which together contributed to the multi-model ensemble developed in the project.

Two streams of coordinated seasonal–decadal experiments were carried out during the project:

Stream1 covered the 1991–2001 hindcast periods for seasonal to annual range with 7-month-long hindcasts started every May and November. The November start dates were extended to 14 months in order to cover a full calendar year. Each of the groups contributing to the multi-model ensemble ran nine member ensembles sampling uncertainties in the observed initial conditions. In addition, further nine-member ensembles were run to assess the stochastic physics and perturbed parameter approaches to sampling modelling uncertainties, using the IFS/HOPE and DePreSys systems, respectively. Papers documenting the results have been written (Berner et al., 2008; Doblas-Reyes et al., 2009). The perturbed parameter hindcasts were also tested in decadal prediction mode by extending the hindcasts for all 22 start dates. Partners contributing to the multi-model ensemble also carried out test decadal projections for two start dates (November 1965 and November 1994), using the results to inform the design of the subsequent stream 2 hindcasts.

Stream2 hindcasts consisted of a comprehensive set of seasonal, annual and decadal integrations. The seasonal (7-month long) and annual (14-month long) hindcasts were performed over the 46-year hindcast period 1960–2005, with start dates every 4 months (February, May, August and November). This gave a total of 184 seasonal hindcasts. Ten multi-model decadal hindcasts were carried out over the same hindcast period, starting every 5 years (1960, 1965, 1970, …, 2005) in November. The 2005 start date also provides a future prediction for 2010–2014. The seasonal–annual hindcasts again consisted of nine ensemble members per model, whereas the decadal runs were done with three members per model. The DePreSys system was used to create a large set of decadal hindcasts, initialised every November during 1960–2005. The ensemble hindcasts consisted of theHadCM3 model variant with standard parameter settings plus eight variants distinguished by multiple parameter perturbations.

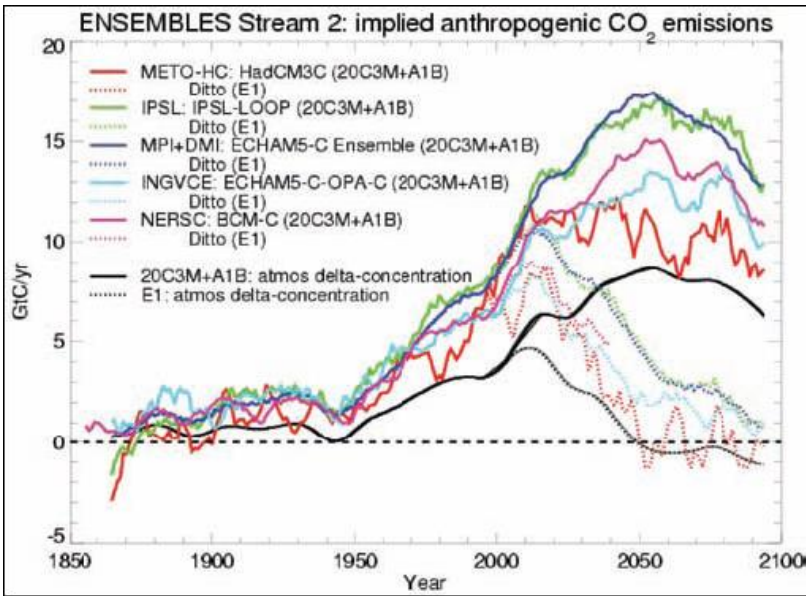
The ENSEMBLES stream 2 decadal hindcasts provided a first opportunity to assess the benefits of combining projections from different models in a coordinated experiment, following initial studies carried out with individual climate models (Smith et al., 2007; Keenlyside et al., 2008; Pohlmann et al., 2009).

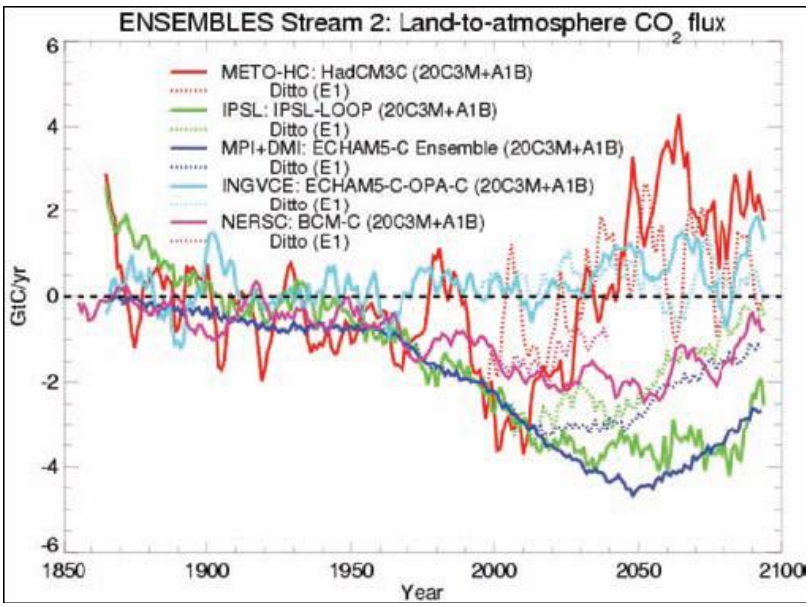
Two streams of coordinated centennial were carried out during the project:

The modelling groups involved (CNRM, DMI, IPSL,METOHC, MPIMET, NERSC and FUB) performed the **Stream 1** simulations using the common set of agreed forcings for the historical simulations over the period 1860–2000, and for the three recommended IPCC SRES scenarios A2, A1B and B1 over the 21st century. Some simulations were extended beyond the year 2100 with constant atmospheric concentrations from the B1 and A1B scenarios. Additional simulations with a 1% increase of CO2 per year with stabilisation at 2×CO2 and 4×CO2 were also performed.

Stream2 simulations make use of improved coupled atmosphere ocean models, as developed in Work Package 1.1 of RT1. A major difference consists in the coupling of an interactive carbon cycle in five Earth system models, so that the net CO2 fluxes between atmosphere and ocean, and atmosphere and land can be computed interactively, depending on the prescribed atmospheric CO2 concentrations, as proposed by Hibbard et al. (2007). The carbon cycle model components generally describe the carbon storage in different pools related to vegetation and soils, and the carbon uptake and cycling in the oceans (see figure 1).

Eight of the models can be forced with land cover changes. Land cover changes influence the physical properties of land surfaces, and imply carbon emissions in models including land carbon pools for vegetation and soils, e.g. when deforestation takes place. Aerosol and/or chemistry models were introduced in three models. Further improvements concerning details of the model formulations have also been included, so that the new set of Earth system models represents a considerable step forward towards future models with prognostic treatment of all major greenhouse gases. Several models used here represent prototypes of Earth system models that will be used for CMIP5 simulations, in support of the Fifth Assessment Report of IPCC. Paper documenting the results has been written (Tim Johns et al., 2011).


ENSEMBLES Stream 2: implied anthropogenic CO2 emissions
METO-HC: HadCM3C (20C3M+A1B)
Ditto (E1)
IPSL: IPSL-LOOP (20C3M+A1B)
Ditto (E1)
MPI+DMI: ECHAM5-C Ensemble (20C3M+A1B)
Ditto (E1)
INGVCE: ECHAM5-C-OPA-C (20C3M+A1B)
Ditto (E1)
NERSC: BCM-C (20C3M+A1B)
Ditto (E1)
20C3M+A1B: atmos delta-concentration
E1: atmos delta-concentration
GtC/yr
Year


ENSEMBLES Stream 2: Land-to-atmosphere CO2 flux
METO-HC: HadCM3C (20C3M+A1B)
Ditto (E1)
IPSL: IPSL-LOOP (20C3M+A1B)
Ditto (E1)
MPI+DMI: ECHAM5-C Ensemble (20C3M+A1B)
Ditto (E1)
INGVCE: ECHAM5-C-OPA-C (20C3M+A1B)
Ditto (E1)
NERSC: BCM-C (20C3M+A1B)
Ditto (E1)
GtC/yr
Year

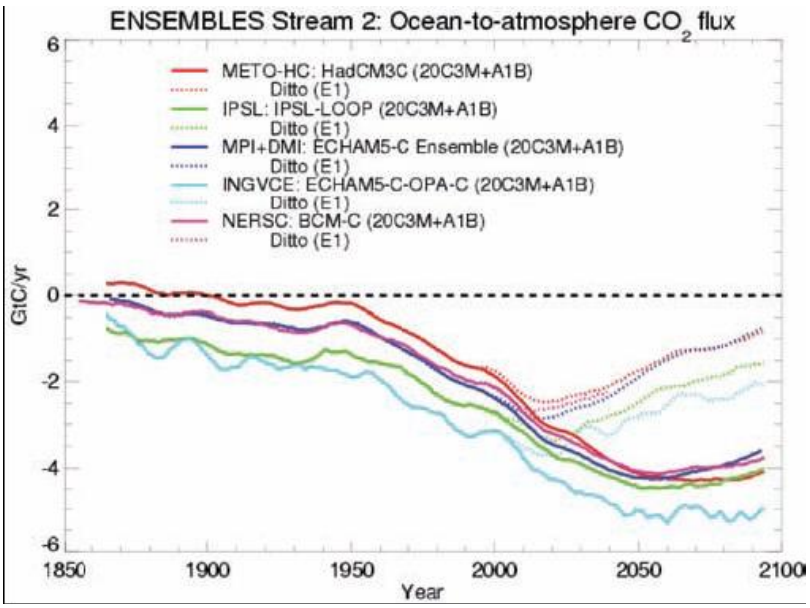

ENSEMBLES Stream 2: Ocean-to-atmosphere CO2 flux
METO-HC: HadCM3C (20C3M+A1B)
Ditto (E1)
IPSL: IPSL-LOOP (20C3M+A1B)
Ditto (E1)
MPI+DMI: ECHAM5-C Ensemble (20C3M+A1B)
Ditto (E1)
INGVCE: ECHAM5-C-OPA-C (20C3M+A1B)
Ditto (E1)
NERSC: BCM-C (20C3M+A1B)
Ditto (E1)
GtC/yr
Year

Figure 1: Implied ('permitted') anthropogenic net carbon dioxide emissions to the atmosphere (Gt C/yr) in ENSEMBLES RT2A Stream 2 runs for the 20th and 21st century (top panel) as diagnosed from the imposed change in atmospheric concentrations (black curves) and the modelled net carbon flux exchange between the atmosphere and the land surface (middle panel) and ocean (bottom panel). Note that an 11-year mean smoothing has been applied to all curves (including delta-concentrations) and that MPI+DMI: ECHAM5-C results show ensemble means of eight (20C3M+A1B) and five (E1) independent simulations, tending to smooth those results compared with other models.

METAFOR Questionnaire

How the METAFOR questionnaire for CMIP5 will be set up to gather information about ENSEMBLES stream1 and stream2 seasonal to decadal and centennial simulations.

A main development in Metafor work package 4 (WP4) is the CMIP5 questionnaire. The CMIP5 questionnaire is a tool for creating CIM documents about the climate models that are contributing towards CMIP5. The CMIP5 questionnaire captures general information about climate model processing- and post processing software, and specific information about how those climate models were configured to perform simulations that conformed to the requirements of the CMIP5 experiments.
A simplified version of the CMIP5 questionnaire will be used to capture metadata for the ENSEMBLES simulations. This will consist of questions about the model software which are less rigorous than those used to describe the CMIP5 models. Specific information about how the models conformed to the requirements of the ENSEMBLES experiments will also be captured.

A model is a software package consisting of at least one or more modules or other software packages. A model includes a set of input conditions which are required to run the model. Running model software with specific conditions is called a simulation. Simulations are made mostly for scientific research purposes and the specific conditions that a simulation must meet are defined by an experiment.

A full description of a simulation consists of information about:
- The model software
  - Version
  - Parameterisations
  - Internal couplings
- How the experiment requirements were met
  - Initial conditions
  - Boundary conditions
  - Spatial temporal constraints
- Temporal and spatial coverage description
- Post processing software
- Other running parameters

The ENSEMBLES questionnaire will capture what is known of the information listed above for each ENSEMBLES simulation.

TDS2CIM tools

How the METAFOR tools (especially tds2cim) can gather information from ENSEMBLES (stream1 and stream2) seasonal to decadal and centennial datasets.

Using models inside experiments produce a large amount of data files. These files are stored in traditional data archives. Several file formats like GRIB, netCDF, HDF or others are used for climate data files. The most common file formats are the netCDF and the GRIB format. The next IPCC Assessment Report will only use netCDF file format.

NetCDF especially in the CF convention contains (use) metadata describing those data which are stored in the file. These metadata are available for automatic capture and can be automatically transferred into the CIM metadata schema.

Regarding ENSEMBLES datasets, the aim was to develop a database system in a common format, allowing easy access by all the partners to selected results of the global ensemble simulations. Typically, an atmosphere–ocean coupled simulation can generate about half a terabyte of data for a 100-year simulation if daily fields are stored. A similar amount is found in ensemble seasonal simulations. Model results are stored in the MARS storage system of ECMWF, and of the Climate and Environmental Retrieval and Archive (CERA) database system at the World Data Centre for Climate hosted by the Data Management Group in Hamburg (DKRZ). Common lists of variables and the need for a common format were outlined in the early stages of the project, depending on the requirements of the scientific community taking part in the other Research Themes. Project data will still be updated and available online after the project has ended.

We have then two different data sources for the purpose of our demonstration.

A large set of atmosphere and ocean variables from the multimodel, stochastic physics and perturbed parameter experiments s2d (seasonal to decadal) integrations are centrally stored at ECMWF for quality control, basic forecast quality assessment, and dissemination. The atmospheric variables are archived on ECMWF's Meteorological Archival and Retrieval System (MARS) in GRIB (gridded binary) format. The fields are stored following a set of atmospheric conventions, based on the experience gained in DEMETER (Palmer et al., 2004) and the operational European multi-model seasonal forecasts. The encoding of the ocean variables is carried out using rules based on newly developed conventions, with storage of CF-compliant NetCDF files into the ECFS server.

A subset of the data is being publicly disseminated. The list of atmospheric variables includes daily data for temperature, wind, humidity and geopotential at four pressure levels and a selection of the most common surface data and fluxes. Monthly mean data are also available. The ocean output includes monthly means of the ocean analyses and forecasts. They comprise 3-D fields (temperature, salinity and velocity) and a limited number of 2-D fields (e.g., sea level, mixed layer depth, 20°Cisothermdepth). For a full list of atmospheric and oceanic variables see:
http://www.ecmwf.int/research/EU_projects/ENSEMBLES/data/common_variables.html.

The ENSEMBLES s2d data have been made available over the internet without charge for use in research, education and commercial work. Two dissemination systems, one based on MARS and another one based on the Open-source Project for a Network Data Access

Protocol (OpenDAP), have been developed to help users to access the ENSEMBLES data in the most efficient way for their specific requirements.

DKRZ Data Management Group has established a website to enable easy access to the ENSEMBLES-related multi-decadal simulations. After providing the necessary information to the data-providing centres, metadata for most experiments have been completed. The web site is continually kept up-to-date to include in the CERA database the datasets provided by the modelling groups. ENSEMBLES data are archived in the Climate and Environmental Retrieval and Archive (CERA) of the World Data Centre System for Climate (WDCC) run by the DKRZ Data Management Group. Access is given by http://ensembles.wdc-climate.de. The experiments for Stream 1 and Stream 2 are briefly described and access is given to each scenario variable of each contributing institute.

There are two ways to produce metadata which shall run into the Metafor repository database (eXist, database as data source for the Metafor portal, see figure 2).

1. Access the ECMWF thredds data server and map the information into the CIM dataObject document
This functionality is an extension of the tds2cim tool. This function parses the ECMWF thredds data server, accesses the file OpenDAP server one by one, retrieves the available information and maps it into the appropriate fields of a CIM dataObjects. The result file is uploaded into the eXist database. This tds2cim tool is implemented in the Metafor portal as function to use as a data ingest source. After registration the registered thredds data server is scanned in the provided time steps and all available data are mapped and uploaded into the repository. The mapping to an existing simulation is done inside the repository by another service.

2. Access the CERA Metadata by a provided XSQL output
This mapping functionality uses the Oracle database XSQL capability to export from CERA metadata RDBMS tables by XSQL into a XML. The result XML is styled by a XSLT style sheet into a CIM dataObject document. The link for doing this is:
http://anticyclone.dkrz.de:8080/xsql/cera_map_cim.xsql?request=dataObject&ac=HADGEM_SRA1B_1_DM_evspsbl
- 'request' means the type of output xml format, here CIM dataObject.
- 'ac' means a CERA dataset acronym which is available from the CERA experiment browser GUI.

'HADGEM_SRA1B_1_DM_evspsbl' is (as an example) an acronym of an ensembles dataset (Please refer to: http://cera-www.dkrz.de/WDCC/ui/BrowseExperiments.jsp).

Upload into the Metafor repository will be done by the upload functionality used for the tds2cim tool. The connection between a simulation and data is made inside the repository (eXist database) by a service (like a cron job). This service is looking for simulations without data and data without simulations (inside the database). If it's possible to connect a simulation to data (e.g. by the DRS: Data Reference Syntax) the connection is made by inserting the simulation uuid into the sourceSimulation field of one or more CIM dataObject.

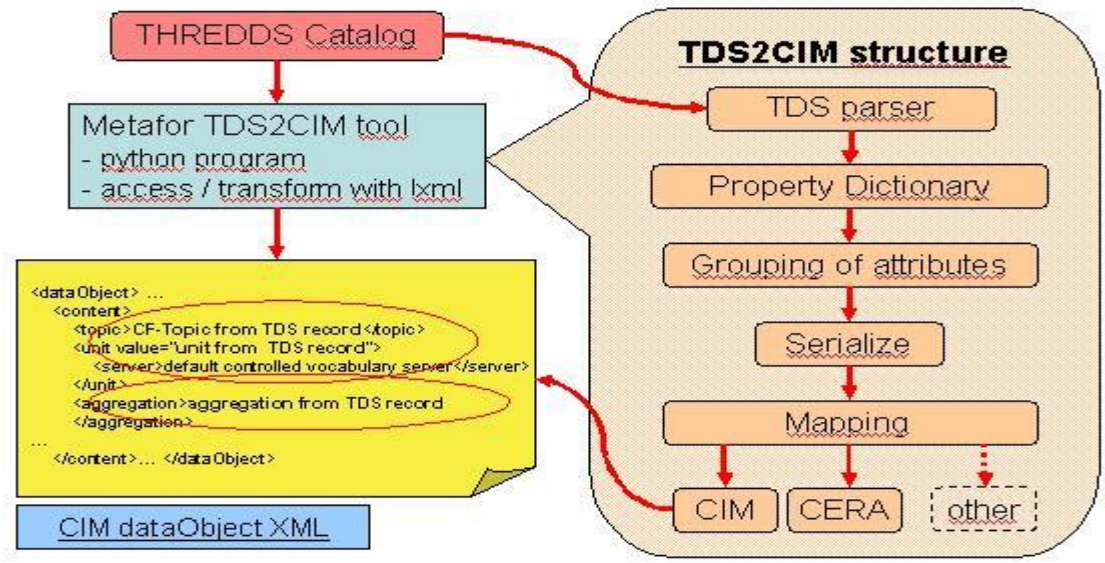


Figure 2:TDS2CIM structure.

Services

How the METAFOR services can be used to ease the discovery and identification of datasets from ENSEMBLES stream1/stream2 seasonal to decadal and centennial datasets.

The Metafor web services have been implemented according to the Service Orientated Architecture (SOA) paradigm. SOA is a key tenet in the design of secure, robust & distributed systems. Such architecture deconstructs a system into a set of discrete functional units known as services. Services are said to be supplied by providers (e.g. Metafor) and consumed by clients (i.e. tools, applications, and in some cases other services).
A web service is such a discrete functional unit deployed upon a web server and thus consumable via the HTTP protocol. Web services leverage inherent HTTP features such as security & caching. By decomposing a system into web-services, rich & diverse informational eco-systems can be incubated, nurtured & supported.
Web services can be implemented in several different styles, but an architectural style known as REST (REpresentational State Transfer) is particularity well suited to the HTTP protocol. REST services posit resources as the unit of exchange between a service & client. A resource is an informational unit supporting at least one representation (i.e. an encoding such as XML).
In our demonstration context REST controllers will be the main resources endpoints. REST controllers will be the METAFOR backend services a third parties portal will query to retrieve information. In particular "Query" as a low level task (see figure3.) will be invoked intensively by the University of Cantabria Portal.

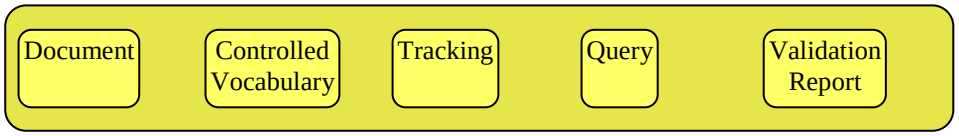


Figure3:
A REST controller manages access to a resource, i.e. an informational unit. REST controllers fulfil their responsibilities by orchestrating calls to the component layer below.

University of Cantabria Downscaling portal integration

How the METAFOR search engine UI will be integrate in the UoC portal.
http://ps.predictia.es/dp

A key aim is to maximize the exploitation of results by linking the outputs of the models results to a range of applications, including agriculture and energy. Thus the UoC Downscaling Portal (http://ps.predictia.es/dp; Cofiño et al., 2007; San-Martín et al., 2009) has been developed following an end-to-end approach to fill the gap between the coarse resolution model outputs and the high-resolution/local needs of end-users. The portal is based on internet and GRID technologies allowing the transparent use of distributed resources, both for data and computation – thus connecting data providers and end-users in a web-based transparent way. The downscaling portal provides user-friendly access to a subset of ENSEMBLES GCM (both seasonal predictions and climate change projections), allowing local interpolation or downscaling to the region/location of interest and bias removal. Users can also upload their own observed grids or networks and interactively downscale the model outputs testing several statistical downscaling methods; including regression, neural networks, analogues and weather typing (see figure 2).

To enhance this experience we plan to include METAFOR search engine UI in the portal to help end-users identifying model results to be used by a large variety of Statistical Downscaling methods.

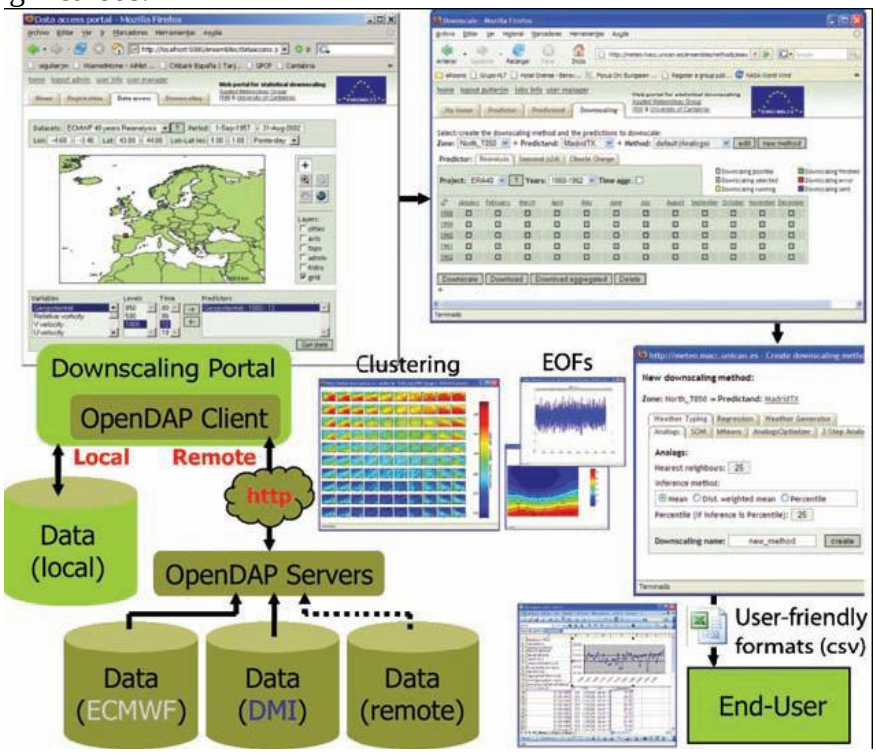


Figure 2: A schematic view of the University of Cantabria downscaling portal.

Conclusions

Due to the complexity of the climate system and the variety of experimental design a climate modeling software can be configured to conform to; it is crucial to offer to end

users a comprehensive representation of this information. The demonstration that has been described here considered end user being outside the climate modelling community. They are not climate modelers but use climate model information to perform their analysis.

This integration describes essential aspects one need to take into account when connecting to the METAFOR CIM eco-system. With respect to this point description of the experimental design and description of the available data are critical point in the demonstration context.

METAFOR provide essential pieces and framework to support parties building climate information system. From the end user perspective (end user from the climate modeling community or from other community) the generic features of both the questionnaire and the tds2cim are essential has they are the major source of CIM information.

Bibliography

[1] ESG publisher scripts: http://www2-pcmdi.llnl.gov/Members/bdrach/.personal/esg-publication-scripts/

[2] Metafor CMIP5 questionnaire: http://q.cmip5.ceda.ac.uk/cmip5/

[3] CERA data model: http://www.mad.zmaw.de/wdc-for-climate/cera-data-model/

[4] World Data Center for Climate (WDCC): http://www.mad.zmaw.de/wdc-for-climate/

[5] CERA database: http://www.mad.zmaw.de/wdc-for-climate/cera-database/

[6] CERA metadata with internal experiment id: http://anticyclone.dkrz.de:8080/xsql/cera_map_cim.xsql?id=2035543

[7] CIM controlled vocabulary, CV: http://metaforclimate.eu/trac/wiki/ticket/192

[8] MindMap CV capturing: http://metaforclimate.eu/trac/wiki/tickets/245

[9] Linux MindMap tool (freemind): http://freemind.sourceforge.net/wiki/index.php/Main_Page

[10] CF (Climate and Forecast) description: http://cf-pcmdi.llnl.gov/

[11] Metafor CIM tools and services: http://metaforclimate.eu/trac/browser/cim_software/trunk/docs/design/METAFOR_D5_5.odt

[12] THREDDS: http://www.unidata.ucar.edu/projects/THREDDS/

[13] DRS: http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf

[14] CMIP5 questionnaire roadmap: http://metaforclimate.eu/trac/attachment/ticket/721/CMIP5QuestionnaireRoadmapBeta5.pdf

[15] TDS2CIM tool / service: http://metaforclimate.eu/trac/wiki/tickets/707

[16] ENSEMBLES final report: http://ensembles--u.metoffice.com/docs/Ensembles_final_report_Nov09.pdf