IS-ENES – WP 10

D 10.1 – Report on CIM dependencies

Abstract:

The Common Information Model (CIM) will provide a metadata framework for the CMIP core archive as well as associated tools for e.g. searching, differencing and browsing. The IS-ENES data portal therefore has to design to be consistent with the Metafor developments. This report summarizes a list of Metafor CIM / IS-ENES dependencies and first steps towards resolving these dependencies.

Additionally first collection steps of a wish list of Climate and Forecast Metadata Convention (CV) variables is summarized, which EU modelling groups would like to have stored.

Grant Agreement Number:	**228203**	**Proposal Number:**		**FP7-INFRA-2008-1.1.2.21**	
Project Acronym:	**IS-ENES**				
Project Co-ordinator:	**Dr. Sylvie JOUSSAUME**				
Document Title:	Report on CIM dependencies		**Deliverable:**	D 10.1	
Document Id N°:		**Version:**	1.0	**Date:**	29/9/2010
Status:					
Filename:	ISENES_Deliverable_10_1_final.pdf				
Project Classification:	Public, Confidential				

Approval Status		
Document Manager	**Verification Authority**	**Project Approval**

Revision Table

Version	Date	Modified Pages	Modified Sections	Comments
0.1	01/04/10	2	2	Initial version including results from Abington meeting
0.2	11/05/10	2	2	Metadata aspects and METAFOR/CIM repository added
0.3	10/09/10	10	2	Integration of review comments from JRA5/WP11 and from project coordinator Correction of spelling errors

Table of Contents

Executive Summary

IS-ENES is the infrastructure project aiming at the development of the virtual Earth System Modelling Centre (vERC), having been launched to support the modelling community studying the impact of the climate change. Pursuing this goal, IS-ENES will disseminate simulation results, model descriptions together with all relevant information pertaining to the IPCC AR5 report. Technical support in terms of hardware, software and data environments will also be provided.

As part of the project, the main service activity SA2 will deliver the modelling results incorporated in the AR5 by providing the support and tools for the data access. The access to the metadata will be provided together with the data directly related to modelling results.

Efficient access to data services will be provided with the JRA4. This includes protocols and software for the direct access to simulation results or their metadata that will grow in the course of the project.

This document expounds the interrelation of the METAFOR and IS-ENES FP7 projects focusing on their mutual linkage - the Common Information Model (CIM). This report is divided into three parts. The first part - introduction - deals with the organization of the document and gives a brief description of its content. The following chapter reviews the developments in the METAFOR project relevant to the IS-ENES project and the requirements of the latter concerning the former. The following chapter elaborates on the dependencies between the two projects and sets the guidelines for their implementation. The final chapter deals with the IS-ENES task to collect additional climate variables, needed by European (regional climate) modellers.

1. INTRODUCTION

One of the main tasks of the IS-ENES project is to disseminate the results of the climate models to a broad range of users. The central role in this endeavor plays the provision of the standardized information concerning all relevant aspects of data management such as metadata generation, access protocols or applied software tools. The standardization is based on the Common Information Model (CIM) currently under development in the scope of the Metafor Project. This creates a close link between these two projects.

A significant aspect of the Earth System Modeling is to produce and publish model documentation. The documentation of the models, integrated within IPCC AR5, is one of the objectives of the IS-ENES project. Within the METAFOR project, this documentation will be collected and translated into metadata by using the CIM. Concise model descriptions based on CIM will lead to a better comprehension and an efficient application of the modeling results. In this sense, the METAFOR provides a mean to catalog data archives offered by the IS-ENES project.

An efficient utilization of the metadata necessitates the availability of advanced software tools for their manipulation. The Metafor project also addresses the development of such tools. They shall be subsequently incorporated in the portal, being developed by IS_ENES. The transition from Metafor to IS-ENES is planned for late 2010 and early 2011.

.

2. Metafor Overview

Climate modeling is a complex process, which requires accurate and complete metadata in order to identify, assess and use the climate data stored in digital repositories. The EU-funded METAFOR (Common **Meta**data **for** Climate Modeling Digital Repositories) project has developed a Common Information Model (CIM) to describe in a standard way climate data and the models and modeling environments that produce this data.

There is a wide degree of variability between different climate models and modeling groups. The METAFOR partners describe the climate modeling process simply as "an activity undertaken using software on computers to produce data." This process has been described as separate UML packages (and, ultimately, XML schemas). This fairly generic structure can be paired with more specific "controlled vocabularies" in order to restrict the range of valid CIM instances. Derived from the UML description a so-called ConCIM (Conceptual CIM) is generated. For the use together with services or tools this is transformed to a AppCIM (Application CIM). The resulting CIM schema is currently available at version 1.4. The Metafor project is intended to end with the final version 2.0.

The following sections describe the available CIM services, as there are CMIP5 questionnaire, CIM query tool with CIM viewer service, CIM differencing tool, THREDDS to CIM tool and the review and completion web framework GeoNetwork.

2.1. CMIP5 questionnaire

The main CIM GUI to fill in model, software, experiment and simulation metadata for CMIP5 (Coupled Model Intercomparison Project Phase 5) / IPCC AR5 will be the CMIP5 questionnaire. The deployment after development was done mainly by the BADC. This tool will be used inside AR5/CMIP5 project to capture the appropriate metadata. The questionnaire captures all CMIP5 relevant information about model, platform, experiment and simulation metadata as well as information about model input data and parties, which are involved in the modeling and data production process. No metadata is exported without a check against a bunch of rules. These rules come from the technical CIM schema site as well as from the modeling site like basic content rules and relations. Basic rules are from the underlying schema which describes which fields are mandatory, which are optional and if enumerations or codelists are used to restrict field values. Additional rules e.g. for the content of fields could be implemented as rules according to the schematron schema (for further information please refer to: http://www.schematron.com/). This allows reuse in other CIM services, which rely also on the schematron system.

The CMIP5 questionnaire the main window (figure 1) shows several data input tabs, as there are: model, platform, experiment, simulation, files, parties, references and grids.

- Model: Inside the model page a model and its components must be described.
- Platform: The platform is the hardware where a model could run.
- Experiment: The experiment page holds the master index of all the simulation runs by available models of the institute.
- Simulation: The simulation page information is about a special simulation, which came out of the model.
- Grids: Grids insert a grid description to the questionnaire grid list.
- Files: Files are used to describe input and boundary files and their content. The data in these physical files are used to force the simulation in an experiment.
- Citation: The ‘citation’ tag defines how experiments and simulations should be referenced. Normally persons and their titles are listed in such a record.

Parties: Information describes a person or an institute, which appears inside the selection list on the ‘model’

page for: act, principal investigator and founder.

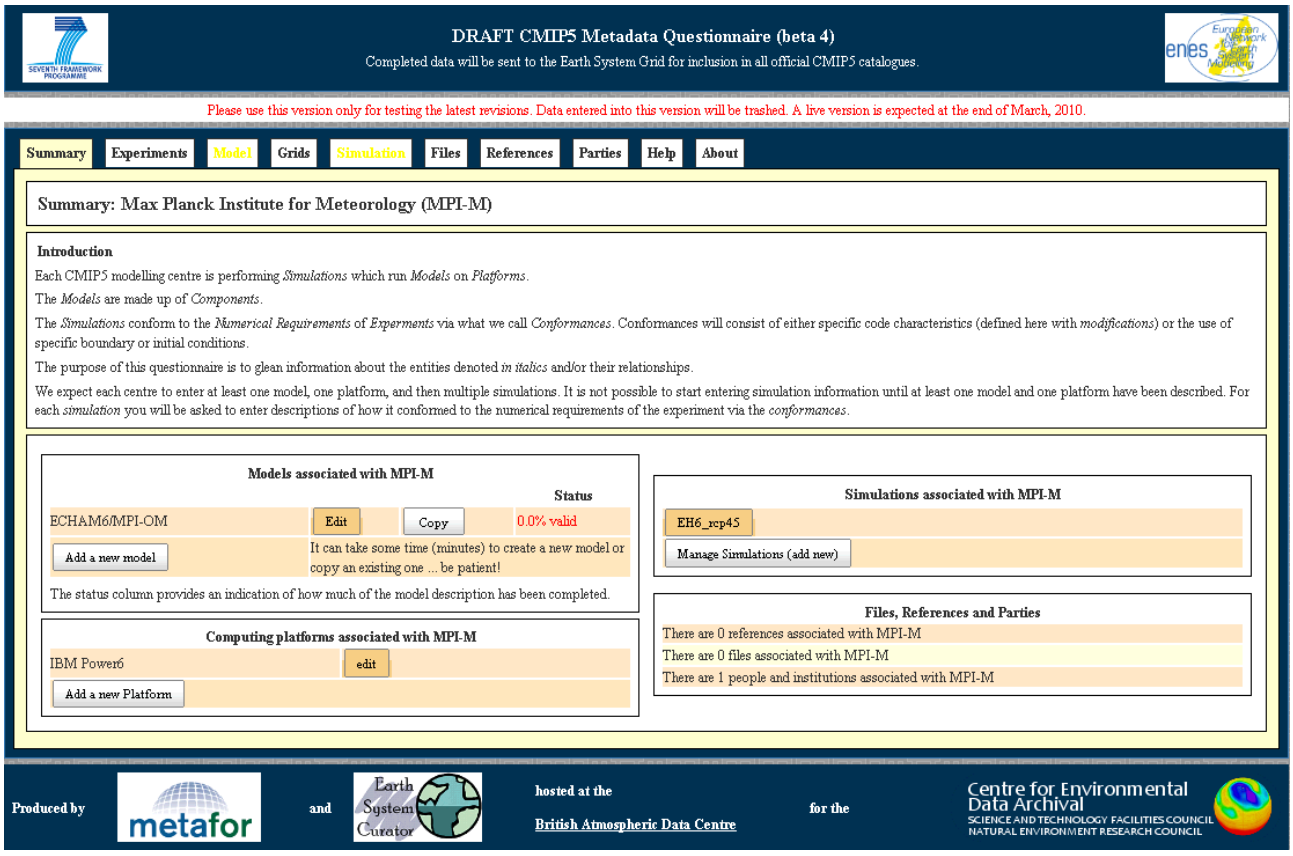

Figure 1 Metafor service CMIP5 questionnaire (http://q.cmip5.ceda.ac.uk/)

2.2. CIM Query and CIM Viewer Service

The Query Tool will allow METAFOR users to perform queries on CIM instances. These instances are stored in a database and accessible via a portal. The Query Tool will allow users to browse and search the portal for the instances they are interested in.

The Query Tool supports three different types of query requests: unrestricted search, advanced search, faceted search. The test implementation is under work. A pilot implementation is envisaged for November 2010.

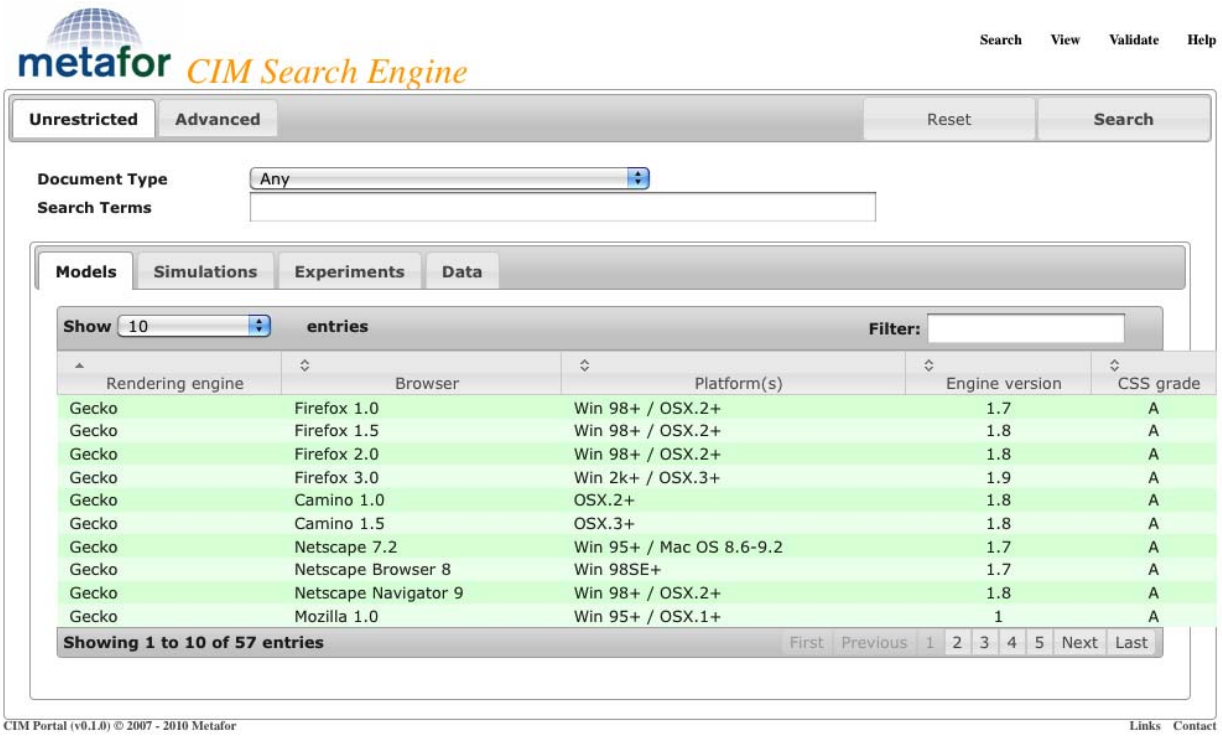

Figure 2 Prototype of Metafor data query service

The query results should be “viewable” and “difference-able.” The former refers to viewing a full CIM instance via an intuitive GUI. The latter refers to viewing the similarities and differences between multiple CIM instances via an intuitive GUI. Initially, the CIM viewer was to be developed as part of the Query Tool itself; however a prototype has already been developed for the CMIP5 Questionnaire. This prototype will be integrated into the final Query Tool deliverable.

2.3. CIM Differencing Service

The Differencing Tool is basically being designed as a feature comparison tool.

Firstly, there are likely to be several variants of comparisons depending upon the type of CIM instances being compared and the type of information being requested.

Secondly, the set of features being compared is potentially orders of magnitude. This large set of features being compared may also change dynamically rather than be hard-coded into the system.

Thirdly, CIM instances have a very rich structure to draw upon. It may be preferable to simply display the presence or absence of a particular feature or to provide the user with a short bit of descriptive text about a particular feature. Or display several detailed and hierarchical points about a particular feature in order to show a user exactly how multiple instances differ.

Clearly, a generic comparison, which ignores the structure of CIM instances and treats each one as a simple text file, would be of dubious value to the end-user. A set of small focused comparisons will be implemented as service accessible from the METAFOR portal. Additionally, these differences are tailored to particular types of CIM documents.

The results of CIM instance comparisons will be displayed as a table with distinct instances on one

dimension and facets being compared on the other. The number of features being compared, and the complexity of those comparisons, will be limited with options widgets.

2.4. CIM Tracking ID Service

The tracking ID service will be implemented to serve several different use cases. The first is to provide a UID to other services from a central UID server. This will guarantee a unique UID inside the CIM world. The second request type is to get a CIM document name or title with a given UID. The third request will deliver a distinct CIM document from the CIM repository or the central CIM database when a valid UID is provided to this service.

2.5. THREDDS2CIM Service

Part of the Earth System Grid (ESG) data node software is the ESG Publisher. It crawls through the file structure (for CMIP5 in Data Reference Syntax (DRS) compliant directory structure) and captures information from the data file headers. The functionality of the publisher e.g. aggregation of data files is controlled by the parameters inside the parameter control file ('esg.ini'). The output runs into the THREDDS catalog, which is accessible via a web browser. All data and possible data services are available through the catalog.

A python script makes the transformation of THREDDS catalog metadata into CIM metadata. This script uses the abilities of the python lxml library (http://codespeak.net/lxml/) to pass the THREDDS metadata into a CIM dataObject template. All unused template metadata will be removed from the template. The connection between CIM objects e.g. parent - child is made with object references by generated ids (by a 'uuid' function or the implemented trackid service). For CMIP5 the connection between a simulation and the appropriate data (CIM dataObject – CIM simulationRun) should be made by the DRS syntax that means as output of the CMIP5 questionnaire a correct named simulation object is available. It will be found e.g. by the CIM xquery service (please refer to Metafor deliverable D5.1).

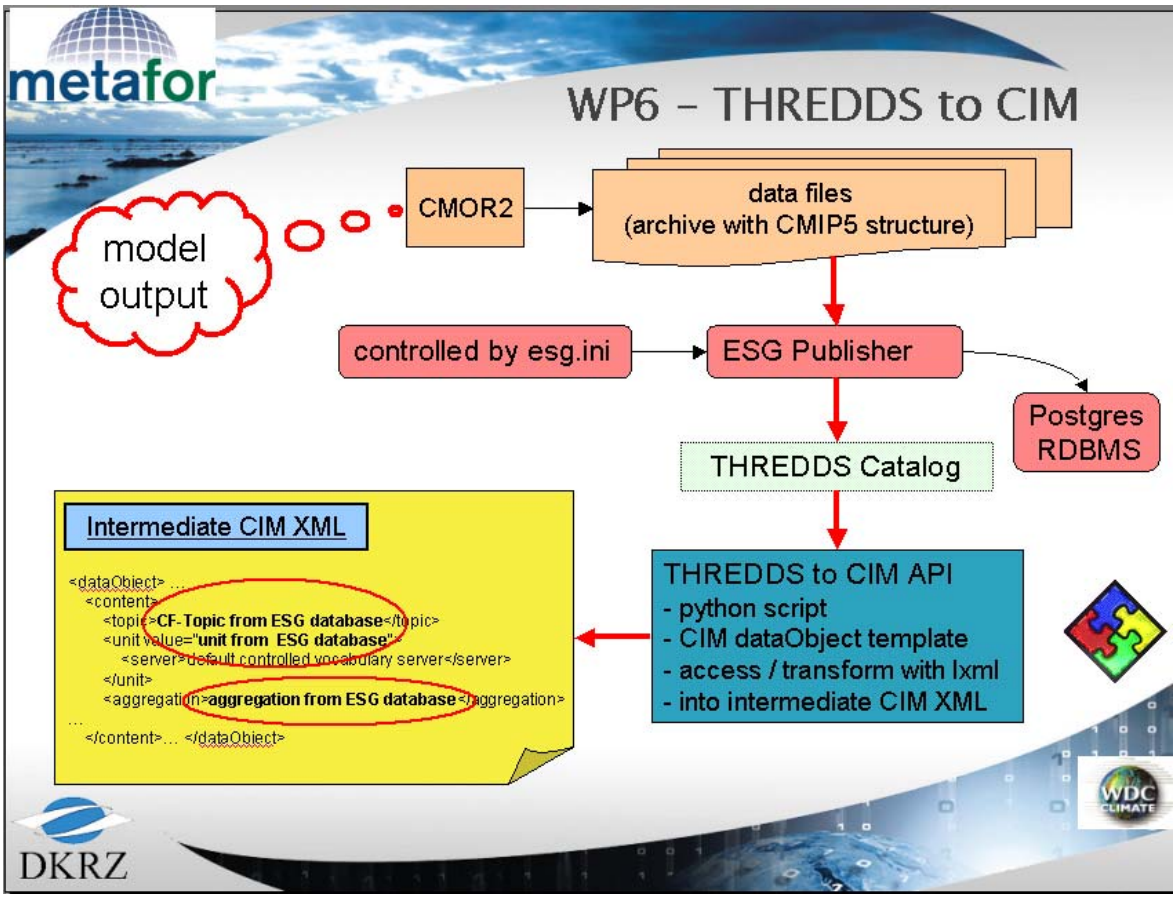

Figure 3 Metafor service, THREDDS catalog metadata to CIM dataObject metadata

2.6. CIM Editing / Completion Service (GeoNetwork)

A GUI for editing and completion of CIM metadata is based on the GeoNetwork framework. To use the CIM in GeoNetwork, the CIM schema was implemented in the GeoNetwork framework including existing schematron rules for validation. With this extension the actual version of GeoNetwork is able to upload, create, edit and validate CIM records. Publishing of the XML records is actual possible via the OAI mechanism. For quality assurance the GeoNetwork service allows for metadata validation and completion before they run into the CIM repository. The next two figures shows screen shots of the actual version of GeoNetwork. The first is the screen with the result list of a free text search with the keyword 'HIGEM'.

Figure 4 Metafor service GeoNetwork, search result list

Figure 3 is a picture of the GeoNetwork screen with the result list of an unrestricted search (keyword is: 'HIGEM'). This feature is used to find stored records in the system for update.

Figure 4 shows a screenshot of a page editing a CIM simulationRun object. Fields outlined in red are defined mandatory in the schema and not yet filled in. The checker inside GeoNetwork marks them red. Storing incomplete records is of course possible, for later completion.

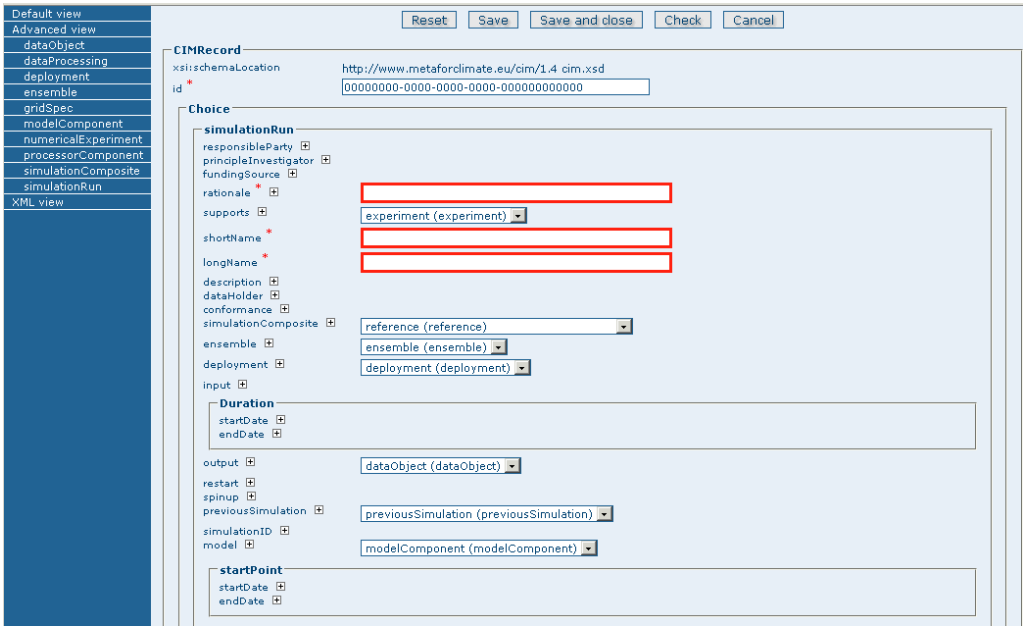

Figure 5 Metafor service GeoNetwork, editing a CIM 'simulationRun' object

2.7. CIM service structure, overview

The next picture shows all available CIM services or tools and how they work together.

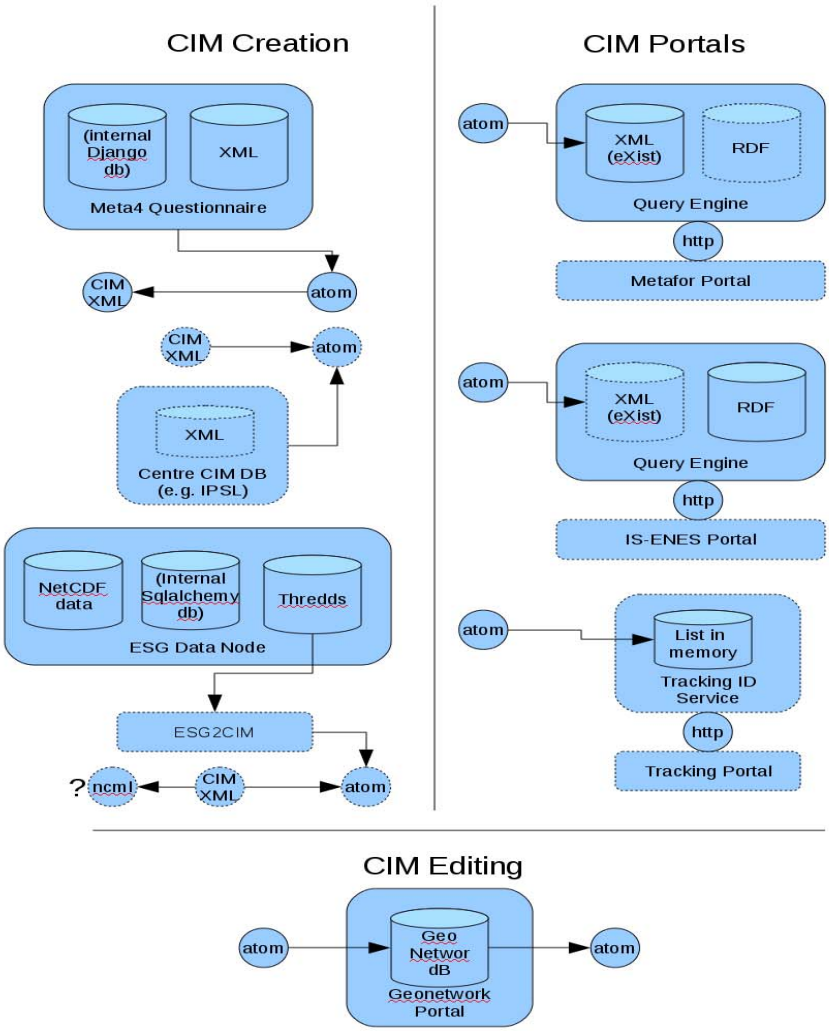

Figure 6 Metafor services and tools structure

2.8. Requirements of the IS-ENES project

The data created within the METAFOR project as well as the tools developed for their access and management are both utilized in various parts of the IS-ENES project:

The vERC portal will use CIM descriptions to provide all relevant information on the earth system modeling to the scientific and modeling community. This information also includes the metadata concerning the results of the calculated models and their descriptions. These, in turn, will be delivered by modeling groups through a

web-based questionnaire. The standardization is a necessary requirement for the effective use and exchange of information.

IS-ENES also has to rely on interfacing tools for the dissemination of metadata related to the climate models. The focus is placed predominately on the tools already developed in the scope of the METAFOR Project and the IS-ENES Project package WP10/JRA4.

The METAFOR/CIM repository which is planned as basis for the vERC (data) portal has to be complete with respect to browse, search and access CMIP5/IPCC-AR5 data and additional data contributions from IS-ENES partners. Additionally the IS-ENES data process workflows (WP10/JRA4 and WP11/JRA5) have to be specified from and controlled by metadata from the METAFOR/CIM repository. The METAFOR CMIP5 questionnaire, which specifies the CIM instances for CMIP5, is planned to be released October 2010. Therefore first CIM repository instances can be expected end of 2010 / begin of 2011. The vERC information portal which links more general information on climate modeling will access the climate model and simulation run information from CIM.

Maintaining seamless link with the METAFOR project is among the objectives of this work package. This will ensure the removal of inconsistencies that may appear in the process. Apart from this, another task requires to move beyond the developments of the METAFOR project and to develop higher-level data services. The emphasis is placed on the management protocols and software providing a distributed archive for the data produced within the 5th IPCC Assessment Report together with these contributed by other European or international earth System Modeling Centers.

2.9. Dependencies IS-ENES - Metafor CIM

An overview of Metafor CIM and (planned) Metafor tools associated with CIM are summarized in section 2.7. CIM service structure, overview

A) an abstract conceptual model formalized in UML (the "CONCIM")

B) an implementation of the abstract model based on XML (the "APPCIM")

C) Tools to handle APPCIM XML instances, including

1) a repository based on an XML database (eXist)

2) a XML instance editor based on Geonetwork portal software (see 2.6)

3) a Web based Questionnaire to collect the metadata and generate APPCIM instances describing models for the Intergovernmental Model Intercomparison Project Phase 5 (CMIP5) (see 2.1.)

4) CIM viewer and CIM query engine (see 2.2.)

5) Atom based feeds to access/harvest/collect CIM instances

6) Automatic generation scripts of CIM based model data metadata instances (interfacing the CMIP5 data node software provided by PCMDI and ESG-CET)

The IS-ENES on one hand relies on CIM to provide model documentation (work packages SA1 and work package 4 NA3). On the other hand IS-ENES interfaces the distributed IPCC AR5 / CMIP5 archive infrastructure (work package 10, JRA4, SA2).

The distributed CMIP5 archive-interfacing requirement introduces an additional (indirect) dependency on CIM: Metafor CIM questionnaire output is currently being extracted and ingested into a CMIP5 gateway installation at NCAR. This includes a translation of Metafor CIM XML into RDF/OWL instances used by the gateway search interface. RDF is used in the future to synchronize the metadata descriptions of distributed ESG gateway installations. IS-ENES has to include and interface gateway installations planed at BADC and

DKRZ in the future and thus has to be able to integrate CIM derived RDF/OWL metadata descriptions and associated tools.

The overlap between METAFOR and IS-ENES was first discussed at the Abingdon workshop 2010-02-24/25. Based on the current status of Metafor and IS-ENES and the knowledge to date the following points were identified:

1. Metafor will release 1.4 CIM XSD at the end of March (with the questionnaire).

2. Metafor xquery software will exploit 1.4 instances. Metafor will not develop RDF tooling beyond helping Curator get RDF from Meta4 1.4 XSD

3. Metafor will leave CIM 2.0 XSD at the end of the project. Design aims for CIM 2.0 should include support for UML->XSD *and* UML->RDFS so that 2.X upgrades can be handled trivially.

4. IS-ENES will investigate RDF tooling, initially exploiting RDF instances harvested via OAI-PMH from ESG gateways.

5. IS-ENES will then migrate to CIM-RDF when and if CIM-2.X-RDF is available (which could be semantically different from the 1.4 being used for ESG and CMIP5).

6. The Metafor portal will concentrate on exploiting xquery tooling for CMIP5 and supporting CMIP5 citation infrastructure. It will be relatively static from early 2011, although it might be able to exploit back ported RDF to add functionality from the IS-ENES development.

7. The IS-ENES portal will continue to evolve, and have enhanced functionality. Backend services may include an evolved version of the Metafor xquery infrastructure consistent with CIM 2.0. Evolving tools to exploit XML instances in CIM2.0 is thus an IS-ENES issue.

8. The IS-ENES portal will interlink Metafor services and provide information on their usage.

2.10. Additional ESM variables from EU modeling groups

The IS-ENES data software suite will depend (as in AR4) on data conforming to the Climate and Forecast (CF) Metadata Convention (http://cf-pcmdi.llnl.gov). Most of the needed CF variables are collected as part of the CMIP5 effort. Additional variables needed by EU modeling groups have to be collected and registered as part of IS-ENES. This section summarizes the status of this effort:

- An initial collection of variables for regional climate model (RCM) forcing for the CCLM and REMO models started (see Appendix B). The collection is based on requirements from DKRZ and is communicated to BADC and IPSL.

- The next step will be to include the requirements from other regional climate modelers contacted by BADC and IPSL.

- Having a first agreed variable collection, the global modelers will be contacted to communicate the requirement for generation and delivery of these variables.

The discussion on the technical details of data production, publication and portal based access is ongoing. There is agreement to follow the data reference syntax (DRS) convention of CMIP5 as close as possible. Nevertheless additional DRS conventions have to be established. The concrete storage format has to be decided as there are requirements to store (parts of) data in native format (not CMOR2-compliant netcdf3).

In addition to the integration of forcing data for regional climate modeling in CORDEX, the integration of CODEX downscaling results themselves is discussed within Metafor and within IS-ENES. It is planned to use the CMIP5 data management structure as far as possible.

APPENDIX A - ACRONYMS

APPCIM – Application CIM

BADC – Brittish Atmospheric Data Center

CIM – Common Information Model

CONCIM – Conceptual CIM

ESG – Earth System Grid

ESG-CET – Earth System Grid Center for Enabling Technologies

NCAR – National Center for Atmospheric Research

OWL – Web Ontology Language

RDF – Resource Description Framework

XSD – XML Schema Document

UML – Unified Modelling Language

APPENDIX B – RCM forcing variables for the CCLM and REMO models

RCM forcing variables from ECHAM6/MPIOM provided in the CMIP5/CMOR2 format
blue variables belong to the CMIP5 list of variables for RCM forcing
Date: October 2009

CF Standard name / CMIP5 standard name	CMIP5 short name	CMIP5 table	ECHAM6 code	ECHAM6 name	Units: ECHAM6 ; CMIP5	Changes every ...	Dimension	Time aggregation ECHAM6	CMIP5 realm / grid	Priority for RCM	CCLM name	REMO name	Data per year[1] for T127 (384,192)
grid_eastward_wind / eastward_wind ??	ua	6hrLev et al.	131	afterburner()	m s-1	6 h	3-D; z-dim=alevel	instantaneous	atmos / atm	required	U		95 * 384 * 192 * 364 4 * 2 = 20.4 GB
grid_northward_wind / northward_wind	va	6hrLev et al.	132	afterburner	m s-1	6 h	3-D; z-dim=alevel	instantaneous	atmos / atm	required	V		
air_temperature	ta	6hrLev et al.	130	stp	K	6 h	3-D; z-dim=alevel	instantaneous	atmos / atm	required	T		
specific_humidity	hus	6hrLev et al.	133	q	kg/kg; 1	6 h	3-D; z-dim=alevel	instantaneous	atmos / atm	required	QV		
mass_fraction_of_cloud_liquid_water_in_air	clw	Amon. cf30min,cfDa	153	xl	kg/kg; 1	6 h	3-D; z-dim=alevel	instantaneous	atmos / atm	required	QC		
surface_air?_pressure	ps	6hrLev et al.	134	aps	Pa	6 h	2-D	instantaneous	atmos / atm	required	P_S		
surface_temperature_of_sea_ice	tsice	Olmon	102	tsi	K	6 h	2-D	instantaneous	sealce / oce	required	TSI		0.215 G
sea_surface_temperature	tos	Omon, da	103	tsw	K	coupling	2-D	instantaneous	ocean, atmos? / oce	required	TSW		
temperature_of_soil_layer / surface_temperature / soil_temperature	tsl? Vorschlag ->KT: 'ts'	Lmon, da, 3hr	139	tslm1 (bare soil & land snow)	K	6 h	2-D	instantaneous	land / atm	required	TSLM1		
sea_ice_area_fraction	sic	Olmon, da	210	seaice	K	coupling	2-D	instantaneous	sealce ocean / oce	required	FR_SEA_ICE		
surface_altitude	orog	fx	129/g	geosp/g	m	once	2-D	constant	atmos / atm	required	HSURF		
land_ice_area_fraction	sftgif	fx	232	glac	[]	once	2-D	constant	land / atm	required			
land_area_fraction	sftlf	fx	194	slf	[]	once	2-D	constant	atmos / atm	required	FR_LAND		
mass_fraction_of_cloud_ice_in_air	cli	Amon, cf30min, cfDa	154	xi	kg/kg	6 h	3-D; z-dim=alevel	instantaneous	atmos / atm	nice to have	QI		
liquid_water_content_of_snow_layer	lwsnl	Limon	141	sn*density	m; kg m-3	6 h		instantaneous	landIce land / atm	nice to have			
surface_snow_thickness	snd	Olmon	214	sni	m	coupling	2-D	instantaneous	sealce / oce	nice to have			
surface_snow_thickness	snd	Limon	141	sn	m	6 h	2-D	instantaneous	landIce land / atm	nice to have			
soil_moisture_content_at_field_capacity	mrsofc	fx	229	wsmax*density	m ; kg m-3	once	2-D	constant	land / atm	nice to have	WSMAX		
soil_moisture_content	mrso	Lmon	140	ws*density	m; kg m-2	6hr	3-D; z-dim=1	instantaneous	land / atm	nice to have	W_SO		
soil_temperature	tsl	Lmon, da, 3hr	207	tsoil	K	6hr	3-D; z-dim=5	instantaneous	land / atm	nice to have	T_SO		
lwe_thickness_of_canopy_water_amount	not required		193	wl	m	6hr	2-D	instantaneous		nice to have	W_L		
sea_ice_thickness	sit	Olmon, da	211	siced	m	coupling	2-D	instantaneous	sealce ocean / oce	nice to have			
[1]at 6hly output intervall; 2byte precision; T127L95													
density= g= all variables are provided every 6hr (if not constant); coupling variables (sni, siced, tsw, seaice) vary according to the coupling interval which is ??? but are sampled every 6 hours; all variables are provided on the atmosphere grid; alevel=model levels		All raw output T127: (104+9 x 95) * 384 * 192 * 364 * 4 * 2 = 206 GB											
		RCM forcing only (required+nice)= (19+ 6 x 95) * 384 * 192 * 364 * 4 * 2 = 126 GB											

Status: final draft