

## METAFOR Support for CMIP5

### METAFOR Deliverable 4.6 M42

PROJECT	
Project acronym	METAFOR
Project full title	Common <u>Metadata</u> <u>for</u> 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	D4.6 Month 42
Deliverable Title	METAFOR Support for CMIP5
Document Identifier	METAFOR-D4.6_M42
Date	September 20, 2011
Work Package	WP4 Services
Authors	BADC
Document Status	Final Version
Document Link	<a href="http://metaforclimate.eu/documents">http://metaforclimate.eu/documents</a>

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

Document History			
Version	Date	Comment	Author/Partner
0.1	September 20, 2011	First Draft	C. Pascoe/BADC, G. Devine/ Reading
0.2	October 03, 2011	Second Draft	C. Pascoe/ BADC, G. Devine/ Reading
0.3	October 05, 2011	Final Version	C. Pascoe/ BADC, G. Devine/ Reading

## Abstract

The CMIP5 questionnaire has been discussed in other Metafor deliverables namely 2.7 and 4.4. The focus of this document is to demonstrate how Metafor is aiding the CMIP5 community by providing access to questionnaire content. The conventional access route to CMIP5 metadata collected by the CMIP5 questionnaire is via the atom feed which broadcast complete CIM documents that are in turn ingested into the Curator Gateway and the Metafor Portal. In addition Metafor is also making interim Metadata available to the community of climate modelers contributing towards CMIP5 via information tables that we are producing for the CMIP5 reports.

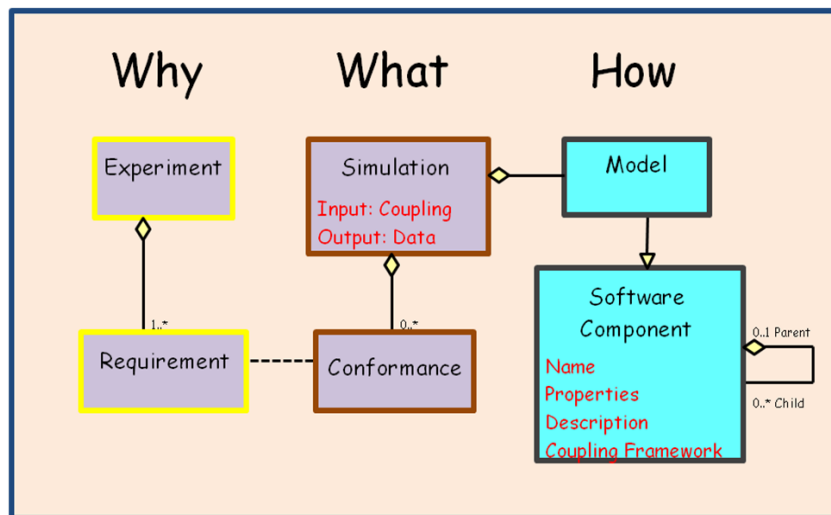
## Table of Contents

<b>ABSTRACT</b> .....	<b>2</b>
<b>TABLE OF CONTENTS</b> .....	<b>2</b>
<b>PURPOSE</b> .....	<b>3</b>
<b>CMIP5 QUESTIONNAIRE METADATA PIPELINE</b> .....	<b>4</b>
<b>ATOM FEED FOR CIM RECORDS CREATED BY THE QUESTIONNAIRE</b> .....	<b>6</b>
<b>COMPLETE CIM DOCUMENTS</b> .....	<b>8</b>
CIM PORTAL .....	8
CURATOR GATEWAY .....	8
<b>INTERIM METADATA</b> .....	<b>9</b>
ENSEMBLE MEMBER INFORMATION. ....	9
IPCC AR5 TABLES .....	10
<i>Progress metrics</i> .....	11
<b>CMIP5 QUESTIONNAIRE HELPDESK SUPPORT</b> .....	<b>12</b>
<b>UPDATES TO THE QUESTIONNAIRE</b> .....	<b>12</b>
<b>CONCLUSIONS</b> .....	<b>12</b>
<b>REFERENCES</b> .....	<b>14</b>

## Purpose

The CMIP5 questionnaire is a metadata infrastructure built to support CMIP5. The contextual information it captures explains why and how climate model data was created from the design of experiments (why) to the implementation of experiments via simulations running models (how) (figure 1). The CMIP5 Questionnaire broadens access to climate model data because for the first time researchers can discover the science encoded in the algorithms of climate models without needing to contact the people who wrote the code. With this contextual information, or metadata, climate scientist are able to analyse more deeply the simulated data produced by different modelling groups.

The CMIP5 questionnaire is based on the Metafor Common Information Model (CIM). The CIM uses UML class diagrams to identify elements that need to be described and the relationships between them. Figure 1 shows a simplified view of the CIM elements that are populated by the CMIP5 questionnaire. The experiments are described as a list of requirements that the simulations must conform to. The simulations run models which are made up of software components and these components can contain child components. The CIM structure is populated using controlled vocabularies which are specific terms, precisely defined that have a common meaning to all climate scientists. Metafor deliverable 2.7 contains further information about the Metafor controlled vocabularies.



**Figure 1:** The CMIP5 questionnaire web tool captures metadata about the life-cycle of a climate simulation. Here we see a UML view of the CIM elements that are populated by the CMIP5 questionnaire; they explain why and how the simulated data was created.

The CMIP5 questionnaire ensures that a standardised set of metadata is collected across the spectrum of climate modelling domains and includes more than 400 scientific properties. A substantial effort was (and still is) required on the part of modelling centres to gather the information about each of the model domains that the CMIP5 questionnaire requires users to describe. Furthermore, we have asked users to make this effort at the same time as they are running their simulations for CMIP5. Indeed, CMIP5 modelling groups initially expressed concern about the unprecedented metadata requirements that they have been obliged to divulge for the CMIP5 questionnaire.

However, now that modelling groups have begun to complete some of their simulations for CMIP5 and are wishing to compare them with simulations from other modelling centres the metadata collected by the CMIP5 questionnaire is proving to be a very valuable resource.

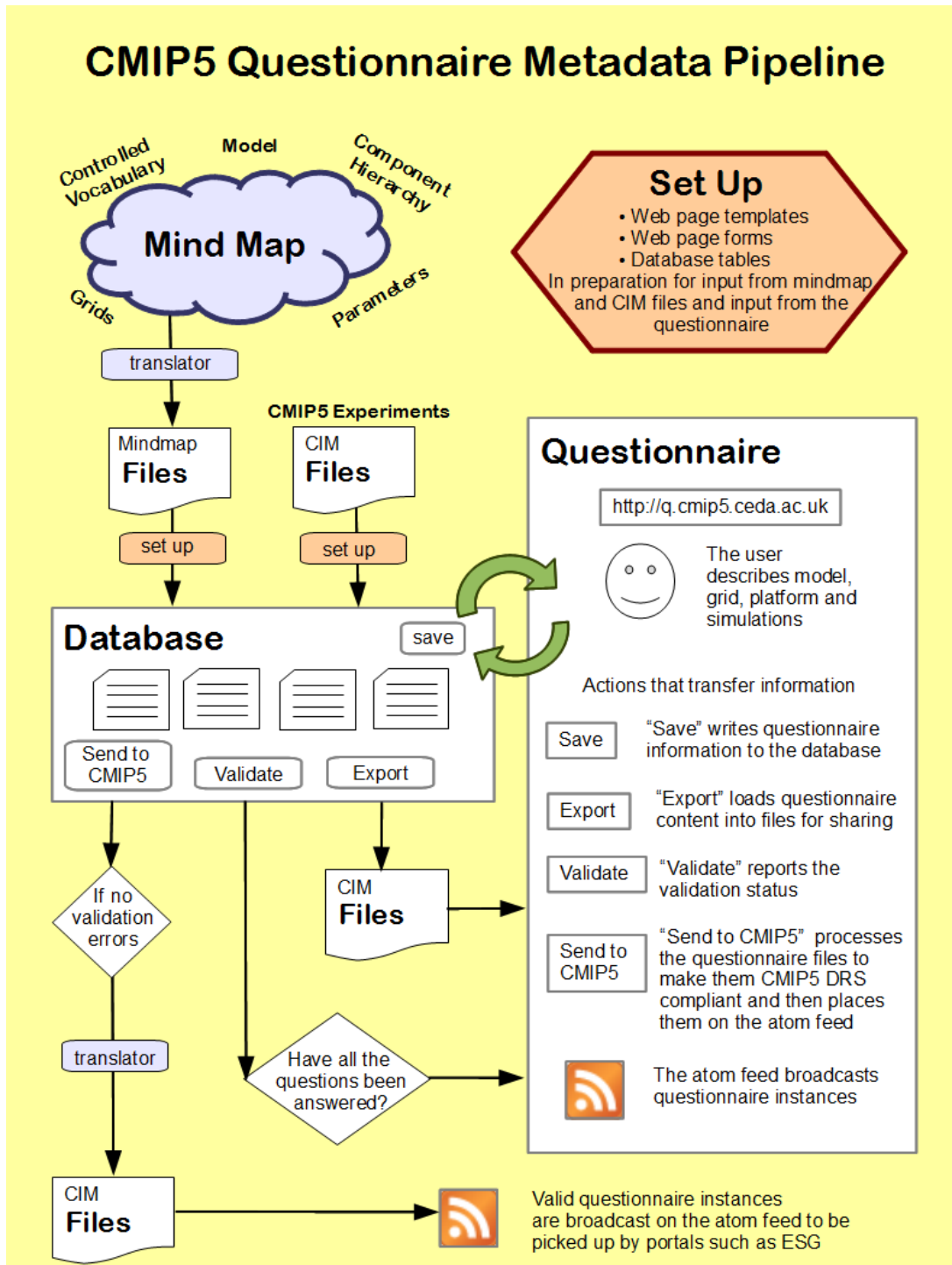
The purpose of this document is therefore to demonstrate how Metafor is supporting the CMIP5 community of climate modelers now by making raw questionnaire data available to them. It will also discuss how the metadata legacy of CIM documents created with the CMIP5 questionnaire will continue to support climate scientist in the coming years.

## **CMIP5 Questionnaire Metadata Pipeline**

The Metafor questionnaire for CMIP5 is an advanced metadata entry system for gathering standardised descriptions of climate models, an overview of the components of the metadata system can be found in figure 2 and is described here.

Mind map software is used to produce xml files for each of the 8 CMIP5 realms (Atmosphere, Aerosols, Atmospheric Chemistry, Ocean, Ocean bio-geo-chemistry, Sea Ice, Land Ice, and Land Surface) that drive the questionnaire. The mind map software is also used to make xml files for other model elements namely the model grid and the computational platform that was used to run the simulations. In addition an xml file for each of the CMIP5 experiments has been hand coded by the Metafor team. The experiment xml files list the requirements of each CMIP5 experiment. The xml files for model, grid, platform and experiments are then processed using python django code to set up a database for the collection of CMIP5 metadata. The django software produces a web interface to the database; it is this web interface that is known as the CMIP5 questionnaire.

The CMIP5 questionnaire uses the mind map structure and controlled vocabularies collected there to create web forms for each model component and sub component of the model realm (Metafor deliverable D 2.7). Users are able to save their metadata descriptions on completing a component page and return again to the questionnaire to complete further metadata descriptions. The Validation function allows users to discover where required information may still need to be completed. Complete and valid descriptions can then be posted to the questionnaire atom feed.



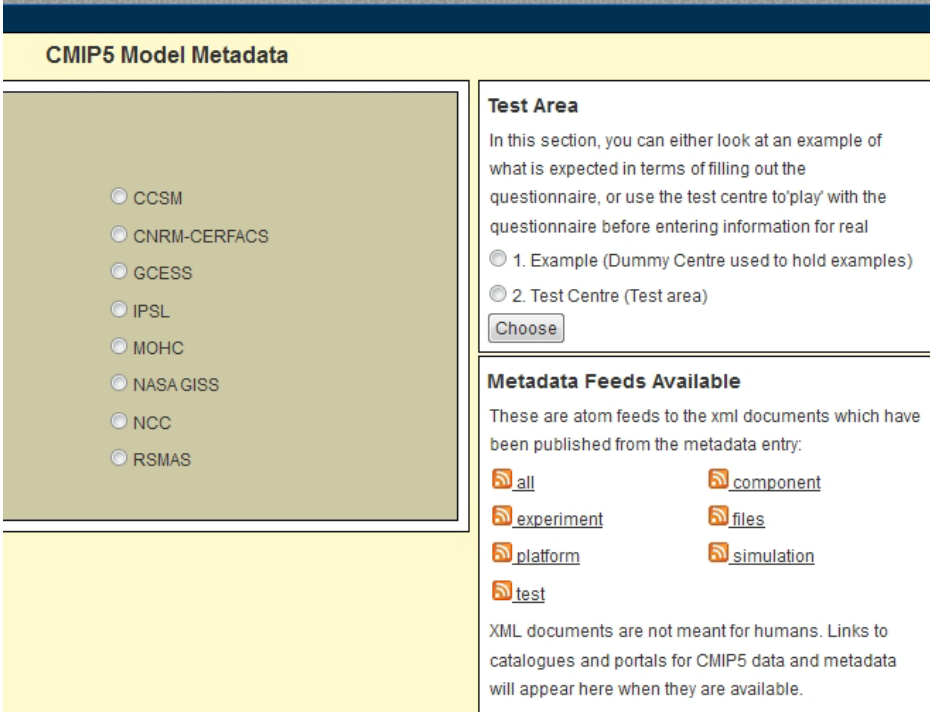
**Figure 2:** CMIP5 questionnaire metadata pipeline. Interviews with climate scientists helped collect basic information needed to understand models, e.g. structured and controlled vocabulary, captured in mind maps. The mind maps together with the CMIP5 protocol description are automatically transformed into a web questionnaire. Once the questionnaire is completed and validated, instances (CIM files in XML), are broadcasted and harvested by several portals (ESG Gateway, Metafor portal, vERC portal<sup>1</sup>), in which the binding with the CMIP5 data files is made.

<sup>1</sup> <https://verc.enes.org/>

## Atom feed for CIM records created by the questionnaire

Atom is a protocol for publishing web-based resources, and, in particular, those resources that are regularly created and/or updated. In this respect, Atom is a suitable infrastructure in which to manage the publication and republication of CIM documents from the CMIP5 questionnaire, and to make these available, via so-called feed readers, to those bodies that consume CIM metadata. In doing so, CIM-aware portals (and other consumers) can automatically inspect, harvest and expose both new and updated CIM documents, on a regular pre-determined basis.

For gaining access to the questionnaire can be found [here](#)



**CMIP5 Model Metadata**

- CCSM
- CNRM-CERFACS
- GCESS
- IPSL
- MOHC
- NASA GISS
- NCC
- RSMAS

**Test Area**

In this section, you can either look at an example of what is expected in terms of filling out the questionnaire, or use the test centre to 'play' with the questionnaire before entering information for real

- 1. Example (Dummy Centre used to hold examples)
- 2. Test Centre (Test area)

**Metadata Feeds Available**

These are atom feeds to the xml documents which have been published from the metadata entry:

- [all](#)
- [component](#)
- [experiment](#)
- [files](#)
- [platform](#)
- [simulation](#)
- [test](#)

XML documents are not meant for humans. Links to catalogues and portals for CMIP5 data and metadata will appear here when they are available.

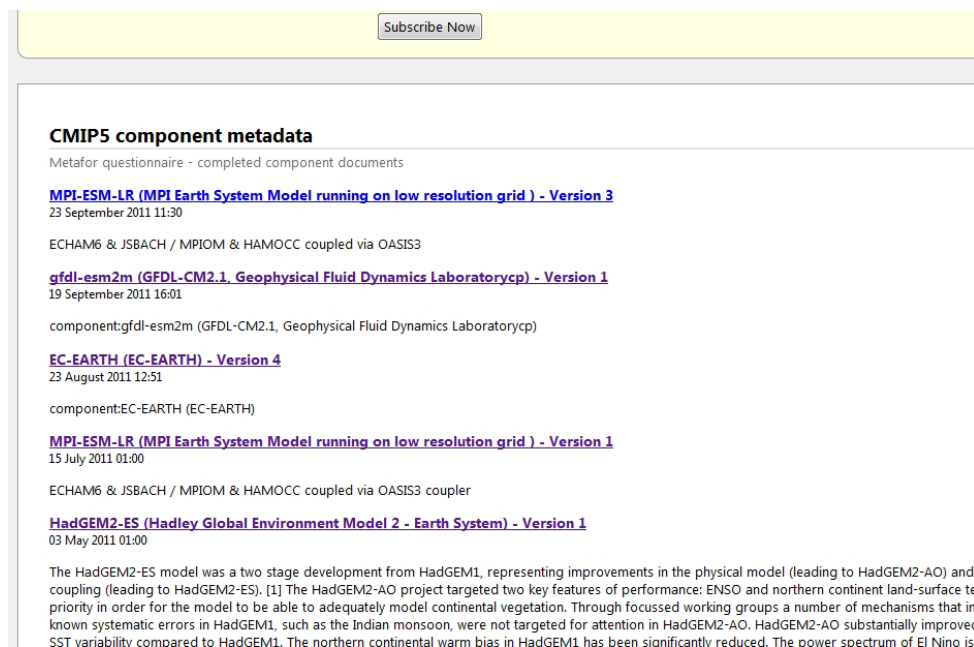
**Figure 3:** Screen shot of the atom feed panel on the CMIP5 questionnaire front page with links to the different types of CIM documents currently in the feed

The CMIP5 questionnaire has the ability to output a number of different individual CIM documents, namely model Component, Platform, Grid, and simulationRun, from the metadata supplied by users to the Questionnaire (other CIM document types are also generated as a subset of a complete simulation, e.g. dataObject). An incomplete CIM document (i.e. CIM-compliant, but incomplete in terms of information required by CMIP5) can be generated at any time during the questionnaire process, which may aid the user in filling out the questionnaire. At the point where a document is fully valid, i.e. it has fulfilled all the requirements necessary to qualify as fully CMIP5-CIM compliant, it can be published to the questionnaire atom feed. This is made possible through a feed-generating mechanism built into Django, the framework on which the questionnaire is built. In doing so, a physical CIM xml document is created and made available to consumers.

Although it is expected that harvesting of these documents will be primarily through automated software, it is possible to manually inspect the questionnaire atom feed using a web browser. The image in figure 3 shows the atom feed panel on the questionnaire

front page with links to the different types of CIM documents currently in the feed (Note that a 'test' feed also exists - this was set up during the alpha release phase of the questionnaire in order to allow upfront testing of the feed by interested parties, and can be used for any further testing of feed documents, whilst remaining isolated from the official feed.)

An actual feed page (figure 4) contains an overview of each of the documents that currently exist in the feed. For each individual item a summary, creation date and version number is given as well as a hyperlinked document title that will link to the actual CIM document itself. This information reflects the underlying representation of each atom feed item, and consequently the information available to feed readers.



**Figure 4:** Screen shot of a CMIP5 questionnaire atom feed page showing model component metadata.

An important aspect of an exported document is the document version number. Upon beginning a new questionnaire instance, a version number is appended to the database information. This version number is then included in the CIM document markup upon publishing. If, following a document being published, the questionnaire information is edited on the same instance, the version number within the database is incremented. When this updated instance is republished a new CIM document will be exported to the feed with a newly incremented version number, alongside the older document. In this way, the full history of documents (whether the most recent or an older version) is available in the feed. This version number can therefore be used to inform feed readers inspecting the CMIP5 atom feed that a new version of this document now exists, and should therefore be harvested.

## Complete CIM documents

### CIM Portal

CIM documents created by the questionnaire and sent to the atom feed are ingested by the Metafor CIM Portal (documented in Metafor deliverable D4.3). The Metafor CIM portal allows for interactive viewing of metadata records (figure 5) in addition to providing search and services functionality for CIM documents. The portal is engineered to ingest CIM records from the Questionnaire atom feed (as well as other CIM sources) on a regular pre-defined basis. Currently this is done on a daily basis. Once ingested, the CIM records are persisted in an eXist xml-database<sup>2</sup>.

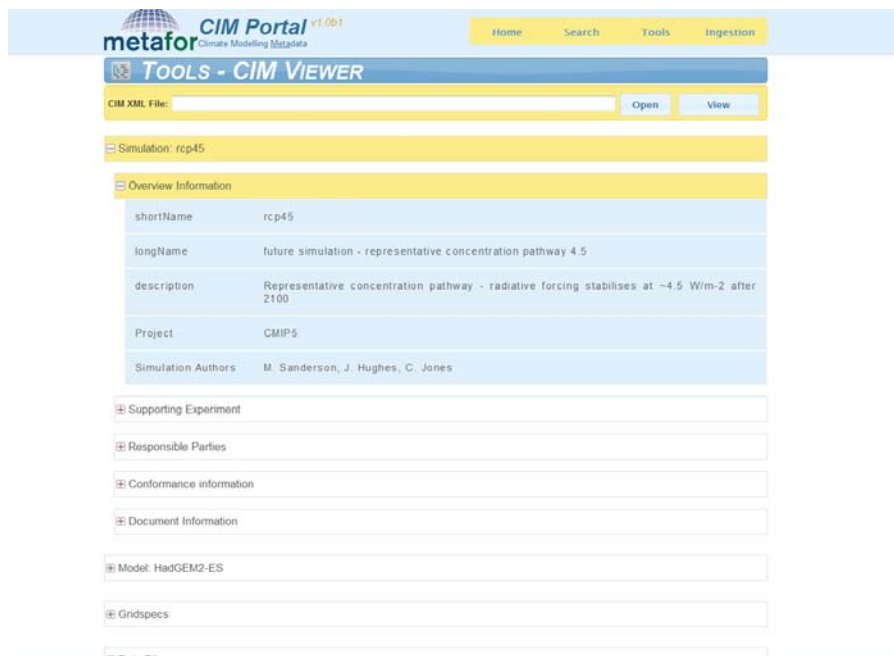


Figure 5: Screen shot of a CMIP5 CIM document viewed using the Metafor CIM Portal.

### Curator Gateway

Curator Earth System Grid [3] provides a gateway to CMIP5 data and also to the metadata collected by the CMIP5 questionnaire. The Curator gateway harvests metadata from the questionnaire via the atom feed and then uses a tool developed by metafor to convert the controlled vocabulary documents into an OWL ontology which is used within the ESGF to support gateway interfaces and guide faceted browsing. This ontology is also used to guide the mapping tool which allows the conversion of CIM documents into gateway triples. These tools (and all Metafor software and schema) can be found on the METAFOR SVN repository at <http://proj.badc.rl.ac.uk/metafor>.

The Curator gateway allows users to search for CMIP5 data by model and experiment name and to discover information about the nature of the experiment requirements and how the simulations conformed to those requirements. The questionnaire support team have been working closely with the curator gateway development team to set priorities

<sup>2</sup> <http://exist.sourceforge.net/>

for the development of the gateway, the aim is to ensure that the information that users discover through the Curator gateway contains the full richness of the information collected by the questionnaire.

For example the CMIP5 experiment known as “historicalMisc” requires the modelling centres to re-run 20<sup>th</sup> century historical simulations but to apply only one of a known list of forcing agents. It is not possible to know from the experiment name alone which forcing agent was used however the list of individual forcing agents are coded into the CMIP5 questionnaire so this information can be retrieved. The questionnaire team have been working with the Curator team to ensure that the metadata elements that contain the forcing information can be found on the gateway interface.

## **Interim metadata**

Information entered into the questionnaire is ultimately output in CIM document format to the questionnaire atom feed. At this point the information is available to the climate community and other interested parties. However, because of the time investment involved in completing complete questionnaire instances, there can be a significant lag time between groups beginning the questionnaire process and having their information output to the feed. In addition, it was decided early in METAFOR that access to information within a different modelling centre should be limited. This therefore puts restrictions on the ability of particular sets of users who need more immediate access to the current questionnaire information across all centers. As part of METAFOR we have addressed two such sets of users and implemented new functionality in the questionnaire to meet their needs.

## **Ensemble member information.**

The CMIP5 questionnaire asks users to describe individual members of an ensemble simulation. In particular it asks for a description of how an ensemble member was implemented, e.g. a change in a parameter value, as well as a ‘rip’ identifier, e.g. r1i1p1, where the r, i and p values correspond to different ensemble techniques as laid out in the CMIP5 Data Reference Syntax (DRS) document [1]. However, individual centers are free to manage their own rip identification and each centre will likely ascribe the rip identifiers in a unique way. The CMIP5 questionnaire is the only place where the modifications associated with rip indices are recorded outside of the data files. Therefore it was deemed necessary by the CMIP5 governing body that metadata collected by the CMIP5 questionnaire about the rip ensemble identifiers should be made available to other modelling groups. The rip information in the CMIP5 questionnaire will allow indices to be understood without the need to download data files and also act as an aid to other groups in the completion of their own ensemble information.

Experiment 1.1 decadal			
Simulation Name:	decadal1965-LR		
Simulation Description:	decadal hindcast experiment		
Number of Ensemble Members	10		
Simulation (first member) rip value:	r1i1p1		
Ensemble Description:	modification of initial conditions		
Ensemble Type:	Initial Condition		
Member 2	INPUT MOD TYPE	INPUT START DATE	INPUT MOD DESCRIPTION
r2i1p1	InitialCondition	1966-01-01T00:00:00Z	Initial conditions taken from a common assimilation run (assim_r2i2p1) for r1i1p2 from 1.1.1966
Member 3	INPUT MOD TYPE	INPUT START DATE	INPUT MOD DESCRIPTION
r3i1p1	InitialCondition	1966-01-01T00:00:00Z	Initial conditions taken from a common assimilation run (assim_r2i2p1) for r3i1p1 from 2.1.1966
Member 4	INPUT MOD TYPE	INPUT START DATE	INPUT MOD DESCRIPTION
r4i1p1	InitialCondition	1966-01-01T00:00:00Z	Initial conditions taken from a common assimilation run (assim_r2i2p1) for r4i1p1 from 3.1.1966
Member 5	INPUT MOD TYPE	INPUT START DATE	INPUT MOD DESCRIPTION
r5i1p1	InitialCondition	1966-01-01T00:00:00Z	Initial conditions taken from a common assimilation run (assim_r2i2p1) for r5i1p1 from 4.1.1966

Figure 6: Screen shot of the ensemble member index page.

Therefore a new feature was added to the questionnaire front page (outside the openID security layer) which collates all such ensemble information into a concise table of information figure 6. In particular, it gives an overview description of the ensemble as a whole, as well as a breakdown of the rip values for each ensemble member and what this rip value refers to in terms of simulation modification e.g. an input modification or parameter change. The new ensemble page also provides users with a dropdown help panel with more detail about how to interpret the information shown in the rip tables.

### IPCC AR5 tables

In order to support the authors of the upcoming IPCC AR5 report, for example in compiling the model evaluation chapter, a further feature is now under development in the questionnaire. This new functionality will generate (from the most up-to-date information entered in the questionnaire) a series of tables that will provide IPCC authors with concise information to be included in the IPCC report. The following sets of information being supplied are:

#### a. Model description table

The model description table collates high-level information about each model documented by the different modelling centers, and includes such information as model grid resolutions, component references, and model vintage. The image below shows a snapshot of the type of information being collated.

MODEL ID, VINTAGE	INSTITUTION	ATMOSPHERE TOP RESOLUTION REFERENCES	OCEAN RESOLUTION Z COORD. TOP BC REFERENCES	SEAICE	COUPLING	LAND
ACCESS1.0_2011	Centre for Australian Weather and Climate Research			rheology: EVP		
CanESM2_2010	Canadian Centre for Climate Modelling and Analysis	top = 0.5 hPa T63L35	256 X 192 depth ,other			
BCC_CSM1.1_2011	Beijing Climate Center, China Meteorological Administration	top = 2,917hPa T42 T42L26	1° with enhanced resolution in the meridional direction in the tropics (1/3° meridional resolution at the equator) Z-coordinate ,linear split-explicit	rheology: EVP		
CMCC-CESM_2009	Centro Euro-Mediterraneo per i Cambiamenti Climatici		2° zonal resolution, meridional resolution varying from 0.5° at the equator to 2° / south of 20°S depth ,linear implicit Madec et al. (1998),	rheology: visco-plastic Fichefet (1997); Fichefet (1999); Timmermann et al (2005);		
CNRM-CM5_2010	Centre National de Recherches Meteorologiques - Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique.	top = 0 hPa none t127r ARPEGE-Climat_V5;		rheology: EVP salas_melia_2002;		Masson et al. 2003; surflex_doc 2009;

**Figure 7:** Screen shot of a model description table. This shows metadata elements that have been entered into the CMIP5 questionnaire which are directly comparable with the model description tables that were produced for the 4th IPCC Assessment report (CMIP3/AR4).

An added feature of this table is the ability to navigate directly from a given piece of information within the table to the questionnaire page where this is entered. Equally, the ability to navigate to the particular questionnaire page from an empty table element is currently being developed. It is hoped that both the upfront exposure of this table to the climate community as well as the added features to navigate directly to particular questionnaire pages will act as extra impetus to the modelling groups to complete their metadata within the questionnaire as well as provide higher quality metadata.

**b. Experiment Description table**

A request was also made to have a concise representation of the particular requirements in place for each cmip5 experiment. The experiment description table will collate the information contained in the experiment documents used to set up the questionnaire database. These experiment documents are an xml encoding of the CMIP5 experiment protocol [2].

**Progress metrics**

It has proved difficult to devise an automated method of determining how far a modelling centre has progressed with their model description part of the CMIP5 questionnaire because there is no obvious relationship between the number of validation errors and the number of components and subcomponents that a centre has yet to describe. However the model description tables offer a useful snapshot of the progress of the modelling centres. Tables such as the model description table in figure 7 can be generated automatically from the questionnaire database and used to target assistance towards groups that have data missing from these tables.

## CMIP5 Questionnaire Helpdesk Support

The CMIP5 Questionnaire is by necessity a complex metadata collection tool and significant effort has been devoted to assisting the CMIP5 modelling groups as they complete the questionnaire. Metafor provided a dedicated support helpdesk (Metafor deliverable D4.4) and has encouraged individual modelling groups to join live online and interactive demonstrations of the questionnaire.

The CMIP5 questionnaire has been online and collecting metadata for 12 months and in that time the questionnaire helpdesk (described in Metafor deliverable D4.4) has answered over 150 queries. The helpdesk will continue to be supported by BADC staff beyond the end of Metafor project for as long as is required by the CMIP5 community.

## Updates to the questionnaire

The questionnaire for CMIP5 has been collecting data for 12 months. Since that time the mind map infrastructure that determines the questions and responses asked of users about their models and simulations has remained unchanged. However as with any prototype, system development continues as bugs are identified and rectified and functionality is improved. When updates need to be made the CMIP5 questionnaire is taken offline so that new code can be incorporated before being re-started. Users are given 24 hours notice of periods of downtime for the questionnaire via a dedicated mailing list [cmip5q@badc.rl.ac.uk](mailto:cmip5q@badc.rl.ac.uk).

## Conclusions

The CMIP5 questionnaire is the first attempt to comprehensively describe the science of an Earth System Model. The CMIP5 questionnaire ensures that a standardised set of metadata is collected across the spectrum of climate modelling domains and includes more than 400 scientific properties. The CIM documents created with the CMIP5 Questionnaire are a unique community resource which will broaden access to climate model data. For the first time researchers will be able to discover the science encoded in the algorithms of climate models without needing to contact the people who wrote the code.

The CMIP5 Questionnaire is by necessity a complex metadata collection tool, and a significant effort has been devoted to running the helpdesk support infrastructure which assists modelling groups as they complete the questionnaire.

In addition to the Metafor project has also collaborated with IPCC authors to release interim metadata from the CMIP5 questionnaire database. Model metadata tables, such as those in figures 5 and 6, have been created at the request of IPCC authors and made available outside of the questionnaire security layer. The model metadata tables are a visceral demonstration of the usefulness of the metadata we are collecting. For example, the information that is collected by the questionnaire about the interpretation of the ensemble indices is not collected by any other parties in CMIP5. Also the visibility of interim metadata allows the different modelling centres to compare the clarity of their descriptions with other groups (live) and to return to the questionnaire and update their own entries where appropriate.

The comprehensive description of CMIP5 earth system models and simulations produced as users complete the CMIP5 questionnaire will allow climate scientists to interpret the simulated data produced by many different modelling groups with unprecedented clarity. Furthermore, the CIM documents created from the CMIP5 questionnaire will remain as a lasting legacy to the climate science community.

## References

- [1] Taylor K. E., V. Balaji, S. Hankin, M. Juckes, B. Lawrence, and S. Pascoe (2011), CMIP5 Data Reference Syntax (DRS) and Controlled Vocabularies [http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5\\_data\\_reference\\_syntax.pdf](http://cmip-pcmdi.llnl.gov/cmip5/docs/cmip5_data_reference_syntax.pdf)
- [2] Taylor K. E., R. J. Stouffer, and G. A. Meehet (2009), A summary of the CMIP5 Experiment Design [http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor\\_CMIP5\\_design.pdf](http://cmip-pcmdi.llnl.gov/cmip5/docs/Taylor_CMIP5_design.pdf)
- [3] Curator Earth System Grid web link: <http://www.earthsystemgrid.org/>