The METAFOR project: preserving data through metadata standards for climate models and simulations

S.A. Callaghan,
NCAS/BADC at STFC
Rutherford Appleton Laboratory
Chilton, Didcot
OXON, OX11 0QX, UK
+44 (0)1235 445770
sarah.callaghan@stfc.ac.uk

A.Treshansky
Coelacanth Consulting Ltd
allynt@coelacanthconsulting.com

M.Moine
CERFACS, France
Marie-Pierre.Moine@cerfacs.fr

E. Guilyardi, NCAS-Climate, Univ. Reading, UK & IPSL/LOCEAN, Paris, France
A. Alias, Meteo-France/CNRM,France
V. Balaji, Princeton University, USA
R. Bojariu, ANM, Romania
A. S. Cofiño, University of Cantabria, Spain
S. Denvil, CNRS/IPSL, France
M. Elkington, Met Office, UK

R. Ford, University of Manchester, UK
M. Kolaninski, Climpact, France
M. Lautenschlager, DKRZ, Germany
B. Lawrence, NCAS/BADC at STFC Rutherford Appleton Laboratory, UK
L. Steenman-Clark, University of Reading, UK
S. Valcke, CERFACS, France

and the
METAFOR project team
http://metaforclimate.eu

ABSTRACT

Climate modeling is a complex process, requiring accurate and complete metadata in order to identify, assess and use climate data stored in digital repositories. The preservation of such data is increasingly important given the development of ever-increasingly complex models to predict the effects of global climate change.

The EU METAFOR project has developed a Common Information Model (CIM) to describe climate data and the models and modelling environments that produce this data. There is a wide degree of variability between different climate models and modelling groups. To accommodate this, the CIM has been designed to be highly generic and flexible, with extensibility built in. METAFOR describes the climate modelling 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.

The CIM will aid digital preservation of climate models as it will provide an accepted standard structure for the model metadata. Tools to write and manage CIM instances, and to allow convenient and powerful searches of CIM databases,. Are also under development. Community buy-in of the CIM has been achieved through a continual process of consultation with the climate modelling community, and through the METAFOR team's development of a questionnaire that will be used to collect the metadata for the Intergovernmental Panel on Climate Change's (IPCC) Coupled Model Intercomparison Project Phase 5 (CMIP5) model runs.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]

General Terms

Algorithms, Management, Measurement, Documentation, Standardization.

Keywords

Metadata, climate modelling.

1. BACKGROUND

1.1 Motivation of the technology and what problems are being solved

Climate science plays an increasingly important role for policy-makers, who are faced with the problem of strategic planning at many levels to address the impacts of climate change. In order to support both basic research and effective strategies to mitigate climate change and, increasingly, deal with its impact on society, a wide range of experts from multiple disciplines need both access to data and advice on the suitability of that data for their purposes. This requires better communication about available climate resources, particularly data from model projections for the next decades and centuries.

The World Climate Research Programme (WCRP) encourages or organises a number of repositories of climate model data, both at the institution and international levels. The value of such repositories is well demonstrated by the large multi-model repository assembled for 4th assessment of the Intergovernmental Panel on Climate Change (IPCC AR4) Working Group I (held at Lawrence Livermore Laboratory in the US, at PCMDI). This unique repository has promoted advances in climate modelling science in an unprecedented way, largely because the data is well-described and easy to find by the climate community (since they contribute to it and are hence well aware of its existence).

Whilst being groundbreaking, the community consensus is that the metadata solution applied in that repository cannot scale in size or complexity to the wide range of existing climate datasets [1,2], nor does it adequately describe the models which generated that data in the first place. Information about climate models is an essential resource for understanding and evaluating climate data, but unfortunately at this time this information is not readily available to end users.

Currently climate repositories are poorly connected, with the result that scientists and policy makers are often not aware of what data is available or from what sources, and even if they are aware, they then have to deal with a variety of institution-dependent data information models (i.e. file formats, metadata structures, documentation methodologies, etc.). As a consequence, comparing and contrasting the information about the data, let alone the data itself, is difficult without significant specific expertise.

Furthermore, because climate models generate such huge amounts of data, exploiting that data is a scientific and technical challenge. The models themselves are complex: each climate model run potentially involves several component models (atmosphere, ocean, sea-ice, vegetation, land ice, ocean biogeochemistry, atmosphere chemistry,) coupled together. Each of those component models can be configured in many different ways, including not only different parameter values but also changes to the source code itself. Component models, or even compositions of component models, can have multiple versions, and individual component models can be coupled together and run in a myriad of different ways. The range of variability is immense and largely under-documented in the output data.

Depending on their needs, a model or data user may want to focus on different aspects of the modelling process. However, there is no standard way of describing climate models and the way they are configured, coupled together, and run. This type of information is essential for making accurate comparisons across datasets, and to prevent misinterpretation or misuse of data. METAFOR aims to fill these gaps, thus helping to increase confidence in climate model data and the use that policy makers, planners, scientists or industry make of that data.

1.2 The target audience

The Common Information Model (CIM) is primarily aimed at climate modelers, as these are the users who are most likely to take advantage of the CIM to document the results of their model runs. However, tools built using the CIM structure to discover and interrogate CIM instances will allow a far wider range of user to access the climate model metadata and data. These users would include local and national governments and policy makers, and academics working in the impacts and adaptation areas of climate change science.

A wide range of commercial organisations are also becoming rapidly interested in climate change issues. Increasingly, these private sector companies need access to primary climate model data to inform decision making in their own domain or that of their clients. The improved access to the climate data repositories hence represents a clear economic opportunity for Europe. This requires that the specific needs of these key stakeholders be taken into account when exposing climate data resources to a wide audience.

Table 1 lists the stakeholders and target audience for the CIM, as identified by METAFOR.

Table 1. CIM stakeholders

Stakeholder/Target Audience	Sector	Level
Academic research	Education	International
Climate impacts academic research	Education	European & international
Planning agencies	Public	European & international
Private companies	Private	European & international

2. TECHNOLOGY

To establish the CIM, METAFOR first considered the metadata methods developed by many groups engaged in similar efforts (for example the US Earth System Curator), explored fragmentation and gaps as well as duplication of information, and reviewed current problems in identifying, accessing or using climate data present in existing repositories.

2.1 Novelty and innovative characteristics

The Common Information Model (CIM) is at the heart of the METAFOR project and has therefore involved all project partners and received significant input from other climate modeling groups in Europe as well as the US.

Climate modeling is a complex process with a wide degree of variability between different models and different modeling groups. To accommodate this, the CIM has been designed to be

highly generic and flexible and is stored in UML[1] as a conceptual model, the CONCIM.

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). Figure 1 shows a high-level overview of these packages which include:

a) activity, the climate modeling simulation/experiments/projects, for example the proposed set of CMIP5 (Climate Model Intercomparison Project) experiments.

b) software, the climate model as well as any analysis programs used, for example fully coupled atmosphere, ocean, chemistry models.

c) data, which may be not only the final climate model data served to the community in data centres but could also include data from different stages of the climate modelling process

d) gridspec, a formal description of the geographic grids modelled by software, required by activities, and mapped to by data

e) reusable elements, like a quality control mechanism, as well as external standards such as ISO standards (especially the GML series) that need to be used.

This fairly generic structure can be paired with more specific "controlled vocabularies" in order to restrict the range of valid CIM instances. For example, the UML allows for a ModelComponent with child ModelComponents; a controlled vocabulary might restrict that pattern to an atmosphere component with a child radiation component (but not, say, a sea-ice component).

The high-level conceptual CIM (CONCIM) has greatly facilitated discussion the result of which, CIM v1.5, is now available (http://metaforclimate.eu/trac/browser/CIM/tags/version-1.5.).

The next stage of development will be creating real CIM instances from a range of climate modelers for many different activities.

2.2 Architecture and workflow

As explained above, the CIM is stored in UML as a conceptual model, the "CONCIM." UML is used because it is implementation-agnostic and its intuitive graphical interface facilitates useful discussion among interested parties. The CONCIM is the normative artifact; it is stored in METAFOR's Subversion repository, and it is what gets modified in response to user needs, and it is what is provided to climate modelers and other interested parties.

METAFOR converts the UML CONCIM into an XML application CIM (APPCIM). This is done by first transforming the UML to XMI[2] (a standard, though very large and unwieldy, XML format for describing UML models). Most modern UML editors can do this automatically. An XSL transformation is then run on the XMI to convert it to a series of XML schema (XSD) files. Together these files define an XML schema that individual CIM XML instances must conform to.

XML is the format that METAFOR has decided to use to store and manipulate CIM instances. We have built up a number of tools which work well with XML: the native-XML eXist database, the bespoke CMIP5 Questionnaire, the GeoNetwork XML editor, etc. However, other groups could decide that a different format is preferable for their application schema. As an example, the US Earth System Grid (ESG) project used a faceted search technology to browse metadata instances. This requires a model written in OWL/RDF[3] instead of XML. They could therefore take the UML CONCIM - a model that both they and METAFOR have broad agreement on - and transform it into an RDF APPCIM.

This relationship is shown in Figure 2.

2.3 The CIM in the existing digital preservation environment

Climate metadata is not a new idea. There are already several different metadata formats being used to describe archived climate datasets. METAFOR has been informed by these existing formats. What METAFOR adds to the mix is metadata about the models and activities that both use and generate that data. The CIM does not replace data which is stored with self-describing metadata files (as with NetCDF, for instance). It exists alongside such formats.

The CIM also provides an intentionally very generic structure with which to describe climate models and data. This allows different user communities to adapt the CIM for their own use. METAFOR has concentrated on the CMIP5 user community, and has developed a set of Controlled Vocabularies (CVs) which provide content for CMIP5 CIM instances. Other groups, though, could replace those CVs with other ones.

2.4 Deployment and feedback

METAFOR develops the CIM using a UML conceptual model of climate models and data. This is transformed into an XML model which is then used for implementation. This approach has been described above.

XML CIM instances can be created and/or edited by hand, by using the GeoNetwork[4] XML editor, or by filling in the CMIP5 online Questionnaire[5]. Once created and validated, a CIM instance is stored in an eXist XML database. The METAFOR portal (written in Pylons) exposes a set of services which operate on instances from the database. Primary among these are querying, differencing, and viewing. The first two services are written using Python and XQuery; the XQuery locates and returns the relevant bits from the eXist database. The CIM viewer is written in Python and Django.

[1] Unified Modelling Language

[2] XML Metadata Interchange

[3] OWL – Web Ontology Language. RDF- Resource Description Framework.

[4] http://geonetwork-opensource.org/

[5] http://q.cmip5.ceda.ac.uk

A separate set of JavaScript functions are being written as a proxy library to interface with those services. The METAFOR portal, simply associates form elements with those JavasScript functions. Using the JavaScript proxy library in this way, METAFOR functionality can easily be added to other portals (without having to modify the look-and-feel of existing webpages). Figure 3 shows a graphical layout of the CIM tools and services feeding into the CIM portal and repository.

In addition to the CIM and the portal/tools being developed by METAFOR, a questionnaire for the CMIP5 community has been deployed. This questionnaire allows CMIP5 users to create CIM instances to accompany the data they are producing for various CMIP5 experiments. The CIM itself - because it is so generic - was unsuitable for providing a template for the type of content that the questionnaire should elicit. Instead a set of mindmaps were developed for different topics in climate modelling.

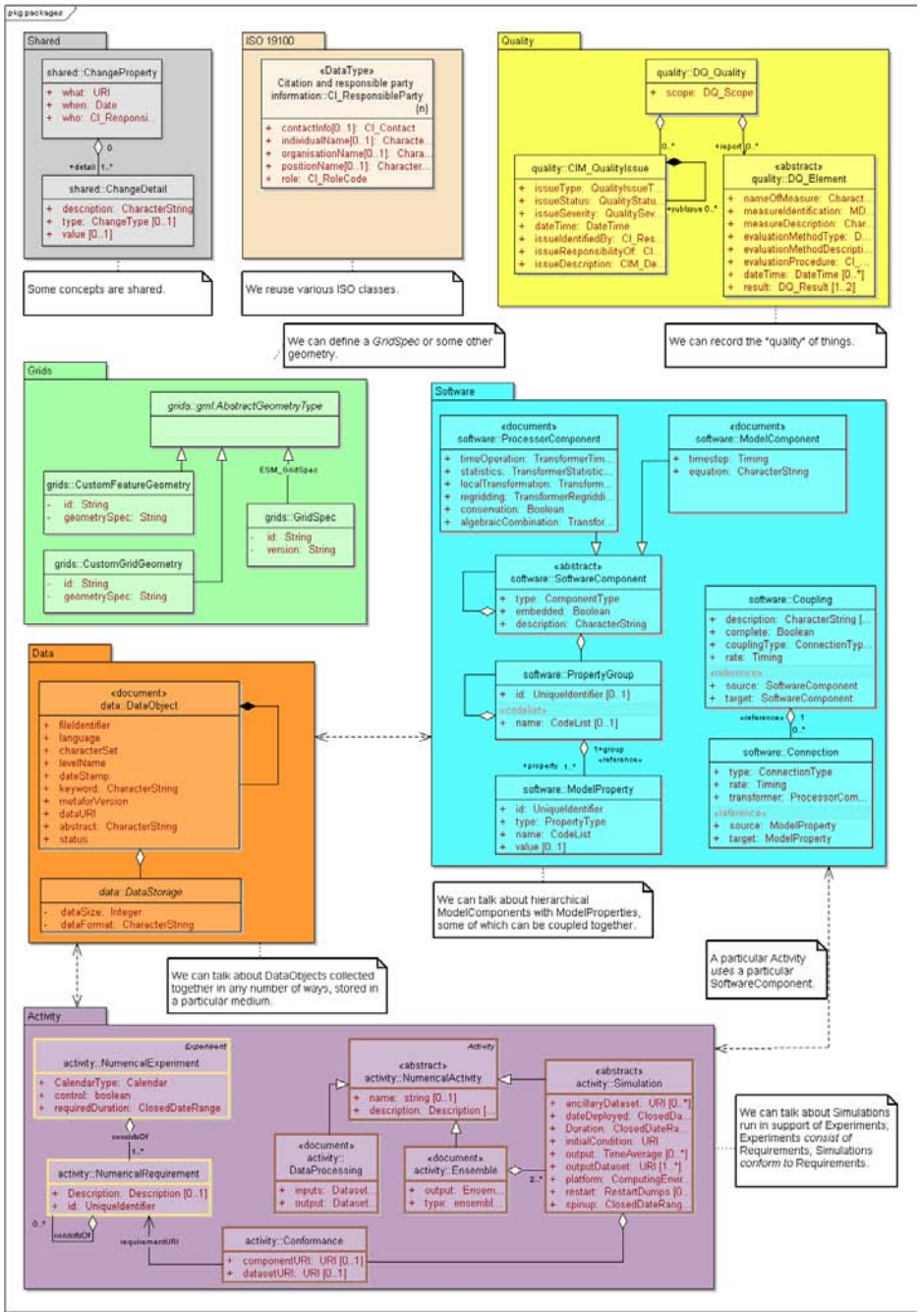

Figure 1 UML overview of the CIM package structure

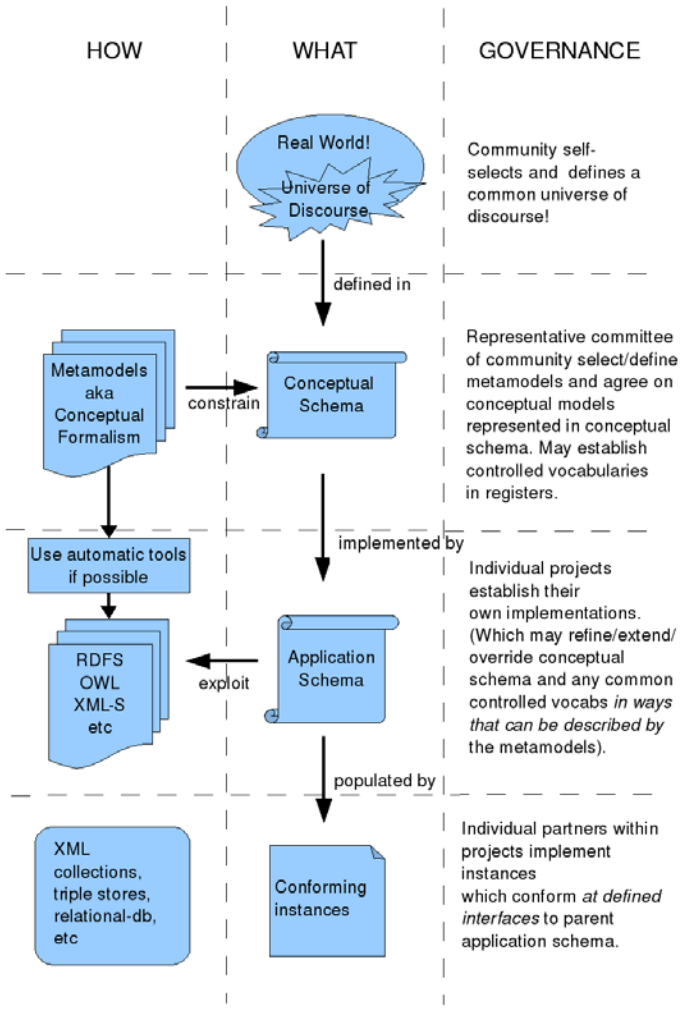

Figure 2 Relationship between CONCIM and APPCIM. There can be multiple instances of APPCIM all related to the same CONCIM, and these different APPCIMs may be implemented in different ways (e.g. XML, OWL/RDF etc.)

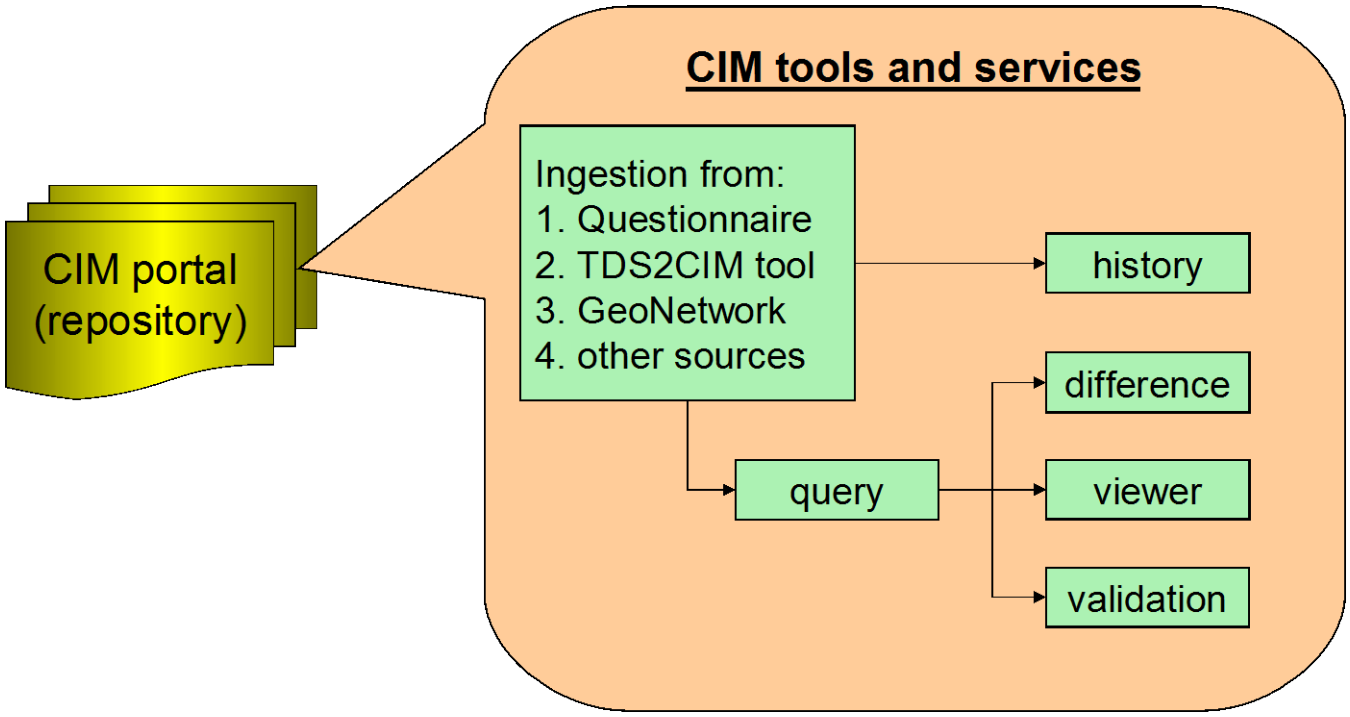

Figure 3. CIM tools and services feeding into the CIM portal.

These mindmaps describe the allowable content of valid CIM instances. The questionnaire uses the mindmaps to configure the set of questions and form elements that are presented to users and that, ultimately, generate CIM instances.

METAFOR spent a great deal of time and effort working with climate scientists - especially those participating in CMIP5 - to create an appropriate set of mindmaps. (In fact, one reason why mindmaps were chosen as a format for storing controlled vocabularies was that it is both visually intuitive and able to be modified in real-time in response to discussions with scientists.) Their feedback, therefore, directly contributed to the content of the questionnaire.

Similarly, the construction of the CONCIM was (and continues to be) developed in consultation with climate scientists and computer scientists involved in climate science. In particular a significant amount of effort was spent getting feedback from the US-based ESG project, which has been developing its own climate metadata portal.

2.5 Documentation, support and community activities

METAFOR has an active mailing list and website which includes the Trac project management and bug/issue tracking system. The site is publicly readable and interested parties outside of the METAFOR project are welcome to join the mailing list. The website also includes a significant amount of formal documentation.

The CIM itself has documentation built into the UML model. This is auto-generated into an RTF file and stored alongside the XSD files comprising the APPCIM.

Finally, there are help files and FAQs being added to the CMIP5 Questionnaire.

The METAFOR team holds weekly teleconferences, where outside participation - notably the US ESG project and the EU IS-ENES project - is welcome.

3. Impact

3.1 Benefits to the digital preservation community

A common metadata standard and a set of tools to locate and analyse metadata documents can help connect producer and consumer. The rich structure of the CIM allows interested users to easily locate the instances they want to review (not to mention the instances related to the instances they want to review). Without something like the CIM, the consumer is forced to consider datasets in isolation from one another and without "provenance" information about how, why, where, when, by whom were they produced. Being noticed is good for the producer of data too - by using the CMIP5 Questionnaire, they ensure that their data is paired with helpful information.

3.2 Productivity enhancement and operational improvement

Creating metadata is an inherently difficult task. METAFOR has improved this process in three ways.

Firstly, for the METAFOR group itself in creating the structure of the CIM schemas: The splitting up of the CIM into a CONCIM and APPCIM has meant that changes to the CIM have been intuitive and straightforward to implement (once the changes have

been debated and agreed upon, of course). Modifying a UML model graphically is much much easier than manipulating an XML schema. Understanding the ideas behind a UML models is much easier than understanding the logic behind a deeply hierarchical XML schema too.

Secondly, for the end-users of metadata METAFOR has created an easy-to-use webform (the CMIP5 Questionnaire) to allow them to easily create and save CIM instances. This is much easier than the alternative of creating an XML file by hand.

Finally, the METAFOR website has provided a central place to store documentation and ongoing discussions about CIM metadata, including recording the progress of the CIM.

3.3 Potential cost saving

Almost all of the tools that METAFOR is using to create the CIM and to create the tools to store and manipulate CIM instances are free and open-source. The tools that are produced by the METAFOR team to edit, manipulate, discover and create CIM instances will be freely available and also open-source, allowing climate scientists to save time and effort by using and modifying these tools rather than developing their own.

4. DEVELOPMENT

4.1 Lessons learned

Building the CIM has benefited heavily from seeking community input. Initial progress was slow as it was largely being designed by computer scientists with an interest in climatology, rather than computer literate climate scientists. Development sped up greatly when METAFOR and ESG began actively collaborating, as each group was able to build on the expertise of the other. METAFOR's relationship with CMIP5 proved another boon; not only did it put us in touch with a new set of climate scientists, it also provided a focused set of use cases (and a strict timetable) to work towards. In retrospect, METAFOR would have benefited by identifying such motivating partners/user groups earlier on in the project.

Maintaining a clear distinction between a conceptual schema and an application schema has been another beneficial methodology employed by METAFOR. It has allowed us to interact closely with scientists, by presenting them intuitive UML diagrams and mindmaps to discuss the domain model we have built up rather than unintuitive and dense XML Schema files.

4.2 Future development plans

METAFOR is currently converting the CIM (v2.0) to a GML-compatible format. Not only will this give us interoperability with other GML technologies, but it will also allow us to use the FullMoon UML to XML conversion tool. METAFOR's bespoke XSL solution for this is rather brittle. The expectation is that FullMoon would come with community expertise and support.

GML domain models also have built-in support for Controlled Vocabularies. Currently, at v1.5 of the CIM, the content of controlled vocabularies is hard-coded into the CIM itself. This is an undesirable feature and should be changed as soon as possible.

The CMIP5 Questionnaire, due to time constraints with CMIP5 users beginning their model runs, will use the version 1.5 of the CIM. Soon CMIP5 instances will start to be saved as users begin setting up their simulations. These instances will be transformed into valid CIM instances and passed on to the METAFOR database. Datasets will not be allowed to be archived at PCMDI as part of CMIP5 without having been first described using the METAFOR CMIP5 Questionnaire.

The METAFOR project finishes in September 2011. It will leave behind the CIM, climate modeling controlled vocabularies, software to manipulate specific versions of the CIM and systems which are populated with instances of specific versions of the CIM. A governance structure is currently being created to manage the continued development of the CIM and the controlled vocabularies. Similarly, an open source community will be brought together to do the same for the software. The METAFOR portal and operational systems will be transitioned to another EU-funded project, IS-ENES[6]. The metadata stored in the CIM database will be maintained for the forseeable future by the BADC[7] as part of their work as a CMIP5 Core Data Node.

5. ACKNOWLEDGMENTS

METAFOR is funded by the EU 7th Framework Programme as an e-infrastructure (project # 211753)

6. REFERENCES

[1] WCRP Informal Report No.8/2008 Report of the 11th Session of the JSC/CLIVAR Working Group on Coupled Modelling (WGCM), 3-5 September 2007, Hamburg, Germany.

[2] D. Bader, "Earth System Grid Development: meeting AR5 needs", at 12th Session of the Working Group on Coupled Modelling, September 22-24 2008, Paris, France

[6] Infrastructure for the European Network for Earth System Modelling https://is.enes.org/

[7] British Atmospheric Data Centre http://badc.nerc.ac.uk