Geosci. Model Dev. Discuss., 5, 1669–1689, 2012
www.geosci-model-dev-discuss.net/5/1669/2012/
doi:10.5194/gmdd-5-1669-2012

Geoscientific Model Development Discussions

Describing Earth System Simulations with the Metafor CIM

B. N. Lawrence[1,2,3], V. Balaji[4], P. Bentley[5], S. Callaghan[2,3], C. DeLuca[6], S. Denvil[7], G. Devine[1,3], M. Elkington[5], R. W. Ford[8], E. Guilyardi[1,3,7], M. Lautenschlager[9], M. Morgan[7], M.-P. Moine[10], S. Murphy[6], C. Pascoe[2,3], H. Ramthun[9], P. Slavin[8], L. Steenman-Clark[1,3], F. Toussaint[9], A. Treshansky[6], and S. Valcke[10]

[1]Department of Meteorology, University of Reading, UK
[2]Centre for Environmental Data Archival, STFC Rutherford Appleton Laboratory, UK
[3]National Centre for Atmospheric Science (NCAS), Natural Environment Research Council, UK
[4]NOAA Geophysical Fluid Dynamics Laboratory and Princeton University, Princeton, NJ, USA
[5]Met Office Hadley Centre, Exeter, UK
[6]NOAA Cooperative Institute for Research in Environmental Sciences, Boulder, CO, USA
[7]CNRS, Institut Pierre Simon Laplace, Paris, France
[8]School of Computer Science, University of Manchester, UK
[9]Deutsches Klimarechenzentrum, Hamburg, Germany
[10]CERFACS: Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique, Toulouse, France

GMDD
5, 1669–1689, 2012

Describing Earth System Simulations with the Metafor CIM

B. N. Lawrence et al.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

Received: 22 May 2012 – Accepted: 22 May 2012 – Published: 22 June 2012

Correspondence to: B. N. Lawrence (b.n.lawrence@reading.ac.uk)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

The Metafor project has developed a Common Information Model (CIM) using the ISO1900 series formalism to describe the sorts of numerical experiments carried out by the earth system modelling community, the models they use, and the simulations that result. Here we describe the mechanism by which the CIM was developed, and its key properties. We introduce the conceptual and application versions and the controlled vocabularies developed in the context of supporting the fifth Coupled Model Intercomparison Project (CMIP5). We describe how the CIM has been used in experiments to describe model coupling properties and describe the near term expected evolution of the CIM.

1 Introduction

Earth system models are used for many things, but two important usages are for providing projections of possible future climate, and for helping advance our fundamental knowledge via contributions to process understanding. These two roles lead to two broad communities of users of earth system modelling: those whose interest is in climate impact and policy, and those whose interest is in the physical earth system itself. While of course there are overlaps between these communities, we can think of these as the climate service community and the earth system modelling-community. Both of these communities require access to data, and crucially, information about that data, to carry out analyses, produce reports, and decide on policy or future scientific experiments. However, the type and detail of the information they require can differ substantially!

Climate data are usually stored in digital repositories, and are sufficiently complex so that accurate and complete metadata (data describing data) is needed for their identification, assessment and use. Each earth system model run potentially involves several component models (e.g. some or all of atmosphere, ocean, sea-ice, vegetation, land

GMDD
5, 1669–1689, 2012

ice, ocean biogeochemistry, atmosphere chemistry, aerosol) coupled together. 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 (at least theoretically). In practice most large models have a number of well understood and extensively tested configurations which have some heritage from previous models. These standard configurations are generally documented in a variety of ways, but often no individual has access to complete documentation for a particular configuration of a model they are running, and it is rare for external (from the modelling group) data users to have access to much documentation for the model and configuration, let alone complete documentation.

Generally the most easily available documentation is found in academic papers, but one finds that to understand a modern earth system model in any detail, one needs access to many published papers, many unpublished papers, and often the personal notes of some key individuals. This can lead to difficulties of scientific interpretation, particularly when comparing the output of two models. For example, when asking questions such as "are the simulation differences due to initial or boundary conditions (and consequential natural variability) or the algorithms/code?". Even with one model it can be difficult to interpret changes between primary configurations (which may differ in ways not being recorded using methods for expediting comparison). Additional complexity arises where models are modified to meet the criteria of a specific experiment (e.g. an experiment to project future climate under a specific emission scenario).

While such difficulties were limited to scientific interpretation, this "model documentation issue, although annoying (and occasionally expensive to work around) was not a major problem. Now that simulations and their validity and uncertainty are the cornerstone of national and international policy, such documentation issues need to be handled differently. To that end, the European Commission established the Metafor project in 2008, aiming to:

> "*... develop a Common Information Model (CIM) to describe climate data and the models that produce it in a standard way, and to ensure the wide*

GMDD
5, 1669–1689, 2012

Describing Earth System Simulations with the Metafor CIM

B. N. Lawrence et al.

adoption of the CIM ... to address the fragmentation and gaps in availability of metadata (data describing data) ... to optimize the way climate data infrastructures are used to store knowledge, thereby adding value to primary research data and information, and providing an essential asset for the numerous stakeholders actively engaged in climate change issues (policy, research, impacts, mitigation, private sector).”

(It is unfortunate that throughout this paper we need to use the word model in two contexts: as something which is used to simulate the real world environment, and as used in CIM, as a construct for describing metadata. Where we use the word model, without qualification, we will mean it in the first context.)

It will be seen that the Metafor project both builds upon, and works closely with other major international efforts, and in particular, the US Curator project (Dunlap et al., 2008).

In 2010, the World Climate Research Programmes’ Working Group for Global Climate Modelling endorsed the use of the CIM, and a questionnaire developed by Metafor, as the mechanism to be used for documenting the models and simulations of the fifth Coupled Model Intercomparison Project (CMIP5, Taylor et al., 2011) .

Along with the motivation of describing models to aid output interpretation, another driver for software documentation is to aid in the construction of models themselves. The coupling of components in an earth system model is often complex, and can involve the moving of fluxes of constituents, energy and momentum from one grid to another, using techniques which need to be very aware of the nature of what is being coupled and how (particularly the source and target grids). Modern couplers are beginning to use automatically generated metadata to aid in that process (e.g. Redler et al., 2010), and it is likely that future models will make even more use of such techniques.

One other possible drive could have been the development of portable and replicable model simulation workflows. While this is in principle true, it is our experience that the portability of current models and production workflows requires significant human interaction, and is likely to do so for at least the next few years. Hence, workflow and

GMDD
5, 1669–1689, 2012

simulation replication is not currently a priority goal for the CIM – although the CIM is being used in workflow experiments to enable provenance description (Turuncoglu et al., 2012).

In this paper we discuss the methodology used to develop the CIM, describe the CIM itself, introduce some of the ecosystem being developed around it, and identify further work. Companion papers discuss the CMIP5 questionnaire (Moine et al., 2012), the application of the CIM in CMIP5 specifically (Guilyardi et al., 2011) and the software infrastructure that supports CMIP5 (Williams et al., 2011).

2 Information context and design methodology

Documentation for climate simulations is not a new idea: the third Coupled Model Intercomparison Project (CMIP3) created an on-line questionnaire to capture key information about the models used, and complex metadata can appear within the data files (e.g. CMIP5 requires file attributes to identify which model was used, and with what forcings key run-time parameters). However, previous efforts have not captured enough quality information to meet the needs of the disparate communities needing simulation documentation. (Nonetheless, where possible, pre-existing concepts from this, and similar, exercises have been co-opted into the CIM.)

Documentation and metadata are also terms which can be misunderstood, so an important decision was to define precisely in what part of the metadata spectrum the CIM was intended to lie. Using the taxonomy of metadata introduced in Lawrence et al. (2009), which describes

- A-Archive metadata (intended to primarily describe the data syntax),
- B-Browse metadata (to provide discrimination between similar datasets, using an inter-disciplinary vocabulary),
- C-Character metadata (for intrinsic quality and extrinsic evaluation, including citation),

GMDD
5, 1669–1689, 2012

- D-Discovery (for location and catalogues), and
- E-Extra (more detailed, disciplinary metadata),

we find the CIM aimed at a mixture of B, C and E applications. Three examples suffice to show the mixture: we want non-physical science climate impacts users to be able to discriminate between simulations (B), we want to be able to record the quality of the archive (are all files present, and have they been checked, are there any citations, C), and we want the CIM to support a detailed scientific comparison of models at the process level and allow software to identify the coupling strategy for software components (E). We assume that discovery (D) is handled independently, with handover from discovery to CIM documentation provided by external software systems.

The difficulty with the requirements of this CIM mixture was that we quickly discovered that there were no pre-existing information structures with rich enough syntactic and/or semantic structures to support our goals, so we needed to develop our own. To that end, we followed the ISO19101 (ISO, 2005a) formalism to identify the key information classes and their attributes, and then built systems around the resulting information objects. This approach requires one to establish formal descriptions of all the important "features" of the domain of interest (in our case, the numerical modelling workflow and all the artifacts used in and/or produced by, such workflows). The resulting set of "feature-types", with their relationships, provides the "domain-model".

ISO19101 recommends the use of the Unified Modelling Language (UML) to develop a domain model encapsulating classes with properties and relationships, followed by the serialisation of that view into an "application schema", typically using the extensible markup language (XML) schema description (XSD). Then, any actual artifacts in the real world (e.g. a simulation or model description), should be described in instances of that schema (i.e. XML documents for an XSD schema).

In practice before using UML in this formalism, one needs to establish a "meta-model" which provides a set of rules that ensures one uses UML in a way that is consistent with the objective (the domain model) and it's eventual serialisation into an

GMDD
5, 1669–1689, 2012

application schema. Between them ISO19101, ISO19109 (ISO, 2005b) and ISO19136 (ISO, 2005c), provide such a set of rules.

Using this method, our implementation was broken into four dependent, but independently evolving, steps:

1. The development of our metamodel – extending the existing default ISO19136 appendix E metamodel to support some specific requirements,

2. The construction of what we came to call the Conceptual CIM, or ConCIM – the UML description of the domain, and

3. The development of our XSD schema implementation of specific versions of the ConCIM (the Application CIM, or ApCIM), and finally

4. The definition of a set of controlled vocabularies that could be exploited within those instances.

By breaking the problem into these four steps, we were able to decouple both the evolution of our understanding of the underlying concepts and the evolution of an implementation of those concepts. This separation of concerns was crucial to our ability to deliver a scientific consensus of how to describe models, a software implementation of the requisite information structures, and tools to use that information. To build later generations of those tools, a fifth step (a serialisation of our ApCIM into JSON via python objects) has also become necessary, but is not discussed here.

In practice the metamodel was developed very quickly, with the main extension from ISO19136 being the use of a “document” stereotype, to indicate that a specific class described a set of information that was intended to have a life cycle of its own – created and managed by different individuals, and perhaps exposed to the internet by services running in disparate locations. This stereotype allowed one to discriminate between objects which might be independently managed and cross-referenced (using the data- or feature- type classes already inherent in ISO19136) but generally under the control of one institution and targeted at one facet of the problem, from those generally controlled elsewhere. An example of the distinction was to discriminate between the classes used to describe an experiment (an instance of which might be defined and exposed by an international body), a model (which might be developed and documented at a particular institution), and a simulation which might be run using that model at a third institution. We also found that a "reference" stereotype was useful in giving clear guidance as to when associations were expected to be serialised by reference to other objects rather than by encapsulating such objects within the object which was the source of the association.

In the next section we describe more fully the ConCIM, concentrating on the key packages and classes.

2.1 The Metafor Conceptual View

Figure 1 shows a high level picture of the key information classes which exist in the ConCIM V1.5 UML, identifying which of them carry the document stereotype, and also identifying the structure by which the classes are organised into most of the key packages. The complete set of packages are:

1. The activity package, which describes the "doing" part of the process by describing data processing and simulations and how those simulations are associated with a specific experimental context (including requirements). Although not shown on this diagram, experiments can then be gathered into projects, such as CMIP5 etc.

2. The software package, which describes the models themselves as well as any analysis or post-processing programs used. The software itself can be decomposed into fully described (and where appropriate, coupled) subcomponents such as atmosphere, ocean etc.

GMDD
5, 1669–1689, 2012

3. The data package, which describes the final data objects produced by simulations and their inputs (including both initial conditions and boundary information used, for example, to force the model with observed data).

4. The grid package (not shown in Fig. 1), which provides formal description of the geographic grids, both when used as computational grids within software components, and those upon which data is projected in data files for input/or output (it is possible that input and output grids may differ from those used internally for computation).

5. A shared package of reusable elements such as customised specialisations of useful ISO classes along with some "orphan" classes such as quality control records and platform descriptions.

The ConCIM is the point of entry for governance of the CIM, being a serialisation independent description of what the CIM should describe, identifying key attributes and their relationships etc. After the end of the Metafor project, the ConCIM is one of the artefacts which will need ongoing governance so that it can evolve as both producer (the modelling community) and user (not just modellers, data users etc) requirements evolve.

In the next section we describe how version 1.5 of this conceptual view has been serialised into a usable XSD application schema. This is the ApCIM version used to support CMIP5, but because of the separation between the ApCIM and the ConCIM, we have also been able to simultaneously learn the lessons of this deployment, and begin work on subsequent versions of the ConCIM. Some of the lessons learned, and their consequences, are described in the final section.

3 The Metafor Application Schema

As described above, the CIM is conceived of, and initially described, in UML, but to use the CIM, one needs something around which tools can be built with which both humans and computers can interact. To that end, the ConCIM can be serialised in a number of distinct ways, here we concentrate on two:

1. as a set of XSD (XML schema documents), one for each package;

2. into the Web Ontology Language (OWL).

There is a semantic mismatch between these two serialisation approaches. In the first case, the clear expectation is that instances are constructed by creating XML documents which conform to the XML schema; in the second, the notion of schema and instances isn't so clearly separated. In the remainder of this section, we will concentrate on the XSD based ApCIM, the OWL serialisation(s) are discussed in the section on creating and manipulating the CIM.

The ConCIM is described using UML, and in particular, using the HollowWorld formalism (https://www.seegrid.csiro.au/wiki/bin/view/AppSchemas/HollowWorld), which amongst other things, imports classes from the ISO series of standards. Provided the UML conforms to ISO19109 and the constraints of the ISO19136 GML rules, the UML can be serialised using such tools into one or more XSD which together make up a GML compliant application schema. A number of such tool chains have been constructed, the two most well known being FullMoon (http://projects.arcs.org.au/trac/fullmoon/) and ShapeChange(http://interactive-instruments.de/index.php?id=28&L=1). We have thus far not used either of the above tools, since when the first version of the ApCIM was developed, the ConCIM did not fully conform to the ISO19136 metamodel. Instead, a completely independent tool was developed using XSL transformations (from an XML representation of the UML) to serialise the UML into XSD. As a consequence, the ApCIM 1.5 is compliant with ISO19109, but not with ISO19136.

Compliance with ISO19136 is not strictly necessary – there are no obvious points of interoperability between CIM documents and complete documents constructed by other communities using ISO19136 – but compliance would make it easier to use tools built by others, and thus avoid having to maintain the entire tool chain ourselves. It

GMDD
5, 1669–1689, 2012

would also help with the translation between CIM descriptions and external data descriptions and discovery systems.

The steps to make future versions of the CIM ISO19136 compliant are not expected to be onerous: the key issues are around adding tagged values which define aspects of the serialisation order, and ensuring that our usage of classes from other ISO standards is done in a consistent manner; in the current version some ad hoc usage patterns have occurred inadvertently.

4 The Metafor controlled vocabulary

The CIM classes introduced earlier define many important attributes, but from the point of view of the users of simulation data, the most important are those which describe the data itself (what is simulated, at what spatial and temporal resolution, and for how long) and the details of the model used. There are already effective metadata standards for describing the data, and so the data package is essentially a wrapper for those. However, the software package is crucial to providing useful descriptions of the models, and within that, their scientific properties, which are related to which algorithm was used, and key configuration parameters. (Another important class of usage, the software configuration properties, describing actual modules of code, which allows, for example, a coupler to join two components together, is discussed in the next section.).

We abstract the scientific properties out of the ConCIM by using two key attributes: model component type and an extensible list of scientific properties expressed as attribute,value pairs. However, the utility of the CIM as an interoperable description of models depends on different groups using these properties in the same way. To that end, we have developed a controlled vocabulary (CV) relating specific components to a set of constrained properties, and the sorts of subcomponents that might be expected. For example, an atmosphere component might expect to have a cloud process sub-component which might have an attribute default particle size, with an expected

GMDD
5, 1669–1689, 2012
Describing Earth System Simulations with the Metafor CIM
B. N. Lawrence et al.

value in m. These controlled vocabularies allow differing software (and algorithms) to be described using a common scientific vocabulary.

The construction of the CV is discussed in Moine et al. (2012), and essentially consists of identifying a set of major model components, along with the hierarchy of sub-components which lie beneath. Then for each component or sub-component, identifying any attributes and/or parameters, and providing formal definitions. These steps were carried out in a series of consultations with many scientists, using mindmaps to mediate the conversations. The mindmaps eventually became the primary artefact not only to record these discussions, but also to serve as the persistent source encoding of the CV. Although their original introduction was because they provided a useful way to develop and display hierarchies and attributes, they are less suitable for machine processing. Nonetheless they became integral to the process because their immediate intuitive use for the scientists was more important than the machine processing issues, for which workarounds were delivered.

5 Using the CIM to control software

The previous section described how the CIM is used to describe the scientific properties of components, and this is where most of the work thus far has been carried out. However, some initial experiments have been carried out using CIM instances to configure the exchanges of coupling data managed by the OASIS coupler within a coupled system. The OASIS coupler (Redler et al., 2010) performs synchronized exchange of coupling fields between component models. In order to do this, formal descriptions are required of the fields to be coupled, the components they belong to, the structure of their grids, the timing of coupling exchanges within the simulation, and any necessary transformations (e.g. temporal averaging or regridding). The CIM can be used to provide these formal descriptions which are then resolved using Connection class instances (see Fig. 2). The OASIS4 coupler was adapted Valcke et al. (2011) so that it could read the CIM XML files containing such instances for its configuration and the

GMDD
5, 1669–1689, 2012

modifications were validated with toy examples testing many features of OASIS4, like regridding, time transformation, bundling of fields and I/O, and debugging output.

6 Creating and manipulating the CIM

The focus of this paper is on describing the construction and structure of the metadata needed to describe earth system models and their simulations; however, metadata is useless without tools to populate and utilise it. Tools to create CIM instances, edit them, aggregate them, and move them into repositories, are needed, as are tools to find specific instances, display and difference them. Prototypes of these tools have been developed. Ideally much CIM content would be automatically created by self-describing models, but as this is not generally the case yet, the construction of a questionnaire suitable for human input has been a major priority. To allow the editing of single CIM instances, particularly those not created via the questionnaire, a customised version of the Geonetwork XML editor (http://geonetwork-opensource.org) has been constructed. A Metafor portal is also under heavy development – it will provide support for a repository of CIM documents with search, view and differencing support, as well as validation and view services for document uploads and general CIM documentation. However, at the time of writing, the most important destination for CIM content is into the Earth System Grid (ESG) gateways described in Williams et al. (2011). An example of a piece of a simulation description is shown in Fig. 3, which represents a view on a number of CIM XML documents. An example snippet of the underlying XML content is shown in Fig. 4 for readers not familiar with XML.

The ESG gateways ingest OWL representations of CIM documents which are created by a tool which effectively maps the ApCIM XSD structure onto a target OWL structure (generated from the ConCIM), and then parses CIM instances to produce triples, which are directly inserted into the ESG gateway triplestores. These then support display and faceted browse of the CIM content (Dunlap et al., 2008). This procedure of conversion is onerous, and it is not resilient to changes in the ApCIM using the

GMDD
5, 1669–1689, 2012

current toolchain, so a new generation of software will shortly be deployed that does not use OWL representations. However, future versions of the CIM which are more consistent with ISO19136 may be able to exploit work underway elsewhere on OWL serialisations of ISO19136 compliant schema and instances.

7 Next steps

While CIM1.5 is implemented to support CMIP5, work has begun on the ConCIM2.0, aimed at addressing three specific syntactic goals: (1) enhancing the metamodel to better support direct serialisation of the model and instances to OWL, (2) refactoring the model so that tools such as FullMoon and ShapeChange can be used to generate XSD without bespoke tooling, and (3) refactoring the model to be consistent with the upcoming ISO19156 Observations and Measurements standard. The first two of these should provide a more transparent mapping between the two existing representations of the ApCIM which should allow, for example, the same portal to easily support faceted browse accompanied with document differencing using the different representations of the same content. The third should allow better metadata interoperation with observational data, and make CIM content more useful in the B- Browse context introduced earlier.

Scientifically, the metadata model is also going to be refactored to address a better separation of concerns between the description of the scientific properties of component models, and their algorithmic implementation. The current version blurs the difference in such a way that a given CIM software instance cannot be used for, for example, both a scientific description using the Metafor CV and the coupling configuration. When resolved, self-describing models will be much more tenable. To this end some early experiments on self-description (and software-metadata consistency) have already been carried out: the Open Fortran Parser (OFP) was modified to output an XML representation of the source code, which was then translated into a CIM document. Clearly not all information is currently explicitly captured in the code, so methods of decorating the

GMDD
5, 1669–1689, 2012

code to add additional information (and/or appropriately configured human interfaces to collect such information at run-time) will be needed.

Further extensions are also needed to support a greater variety of coupling frameworks. The aforementioned syntactic and scientific goals should lead to both wider adoption of the CIM in the earth system modelling community and for documenting environmental simulation software in general. This will be further enhanced by improvements in the Metafor CV: the existing CV was developed with the earth system models of CMIP5 as the main target, with the full IPCC process in mind, work is already underway to extend the CV to support describing downscaling methods and documentation of Impact and Assessment Models. Clearly of course, much work will also continue on developing the tools which generate and exploit the CIM descriptions!

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

References

Dunlap, R., Mark, L., Rugaber, S., Balaji, V., Chastang, J., Cinquini, L., DeLuca, C., Middleton, D., and Murphy, S.: Earth system curator: metadata infrastructure for climate modeling, Earth Sci. Informat. 1, 131–149, doi:10.1007/s12145-008-0016-1, 2008. 1673, 1682

Guilyardi, E., Balaji, V., Callaghan, S., DeLuca, C., Devine, G., Denvil, S., Ford, R., Pascoe, C., Lautenschlager, M., Lawrence, B. N., Steenman-Clark, L., and Valcke, S.: The CMIP5 model and simulation documentation: a new standard for climate modelling metadata, CIIVAR Exchanges, 16, 42–46, 2011. 1674

ISO: ISO19101: Geographic information – Reference Model., Tech. rep., International Standards Organisation, Geneva, 2005a. 1675

ISO: ISO19109: Geographic information – Rules for application schema, Tech. rep., International Standards Organisation, Geneva, 2005b. 1676

ISO: ISO19136: Geographic information – Geography Markup Language (GML), Tech. rep., International Standards Organisation, Geneva, 2005c. 1676

GMDD

5, 1669–1689, 2012

Describing Earth System Simulations with the Metafor CIM

B. N. Lawrence et al.

Lawrence, B., Lowry, R., Miller, P., Snaith, H., and Woolf, A.: Information in environmental data grids, Philos. Transactions of the Royal Society A: Mathematical, Phys. Eng. Sci., 367, 1003–1014, doi:10.1098/rsta.2008.0237, http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.2%008.0237, 2009. 1674

Moine, M., Pascoe, C., Alias, A., Balaji, V., Bentley, P., Devine, G., Ford, R. W., Guilyardi, E., Lawrence, B. N., and Valcke, S.: Development and exploitation of a controlled vocabulary in support of climate modelling, Geosci. Model. Dev. Discuss., in preparation, 2012. 1674, 1681

Redler, R., Valcke, S., and Ritzdorf, H.: OASIS4 a coupling software for next generation earth system modelling, Geosci. Model Dev., 3, 87–104, doi:10.5194/gmd-3-87-2010, 2010. 1673, 1681

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, B. Am. Meteorol. Soc., doi:10.1175/BAMS-D-11-00094.1, 2011. 1673

Turuncoglu, U. U., Dalfes, N., Murphy, S., and DeLuca, C.: Toward self-describing and workflow integrated Earth system models: A coupled atmosphere-ocean modeling system application, Environ. Modell. Softw., doi:10.1016/j.envsoft.2012.02.013, http://www.sciencedirect.com/science/article/pii/S1364815212000618, 2012. 1674

Valcke, S., Epitalon, J. M., and Moine, M. P.: CIM-enabled OASIS, Tech. Rep. TR/CMGC/11/59, http://pantar.cerfacs.fr/globc/publication/technicalreport/20%11/METAFOR_D5.7.pdf, 2011. 1681

Williams, D. N., Lawrence, B. N., Lautenschlager, M., Middleton, D., and Balaji, V.: The Earth System Grid Federation: Delivering globally accessible petascale data for CMIP5, in: Proceedings of the 32nd Asia-Pacific Advanced Network Meeting, 121–130, New Delhi, doi:10.7125/APAN.32.15, http://usymposia.upm.my/index.php/APAN_Proceedings/32nd_APAN/%paper/view/155, 2011. 1674, 1682

GMDD

5, 1669–1689, 2012

Describing Earth System Simulations with the Metafor CIM

B. N. Lawrence et al.

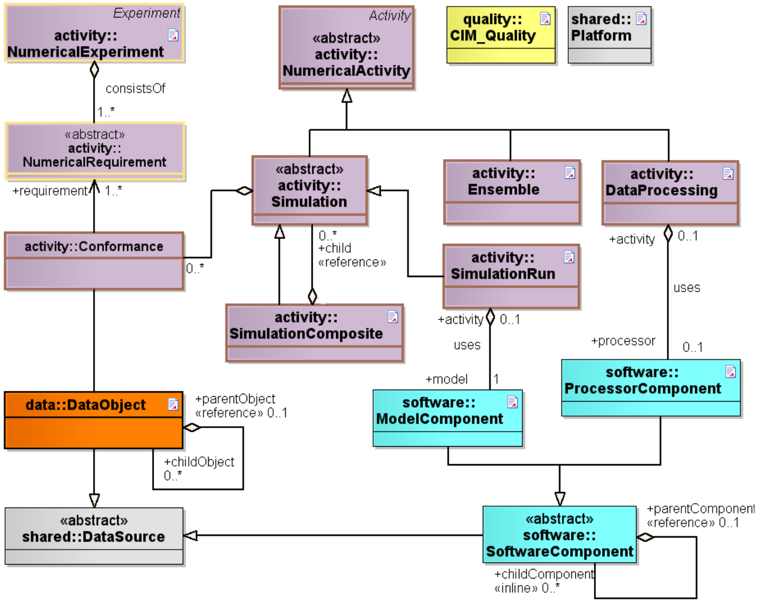

Fig. 1. Key components of the ConCIM V1.5: A SimulationRun is a specific type of Simulation, itself a type of NumericalActivity. A Simulation runs on a Platform, using a ModelComponent, which is a type of SoftwareComponent. A SimulationRun may aggregate SimulationComposites. When components are coupled together, SoftwareComponents can be DataSources for other components (as are data objects from disk). A Simulation will conform to the NumericalRequirements of a NumericalExperiment. All of the entities expected to be managed independently as documents are marked by the document icon in the top right of the box; so we see here that the documents are NumericalExperiments, QualityRecords, DataObjects, two types of SoftwareComponent, and four types of NumericalActivity. (Note that most classes have multiple attributes which have been omitted here to highlight the relationships.)

GMDD
5, 1669–1689, 2012

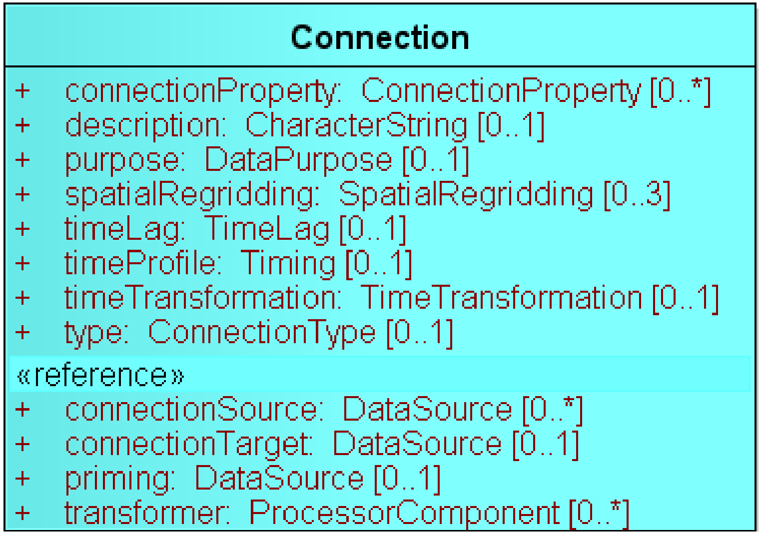

Fig. 2. The Connection class in ConCIM 1.5 which describes the properties of a configured connection between two components.

GMDD
5, 1669–1689, 2012
Describing Earth System Simulations with the Metafor CIM
B. N. Lawrence et al.

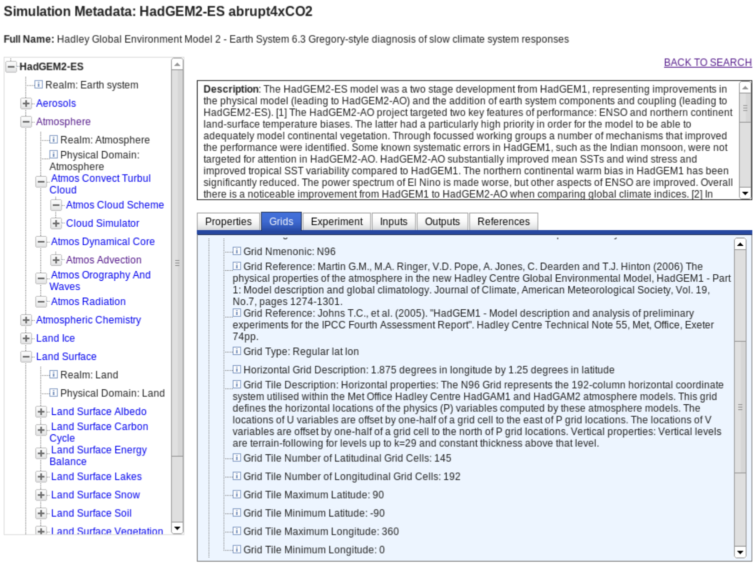

Fig. 3. An example of CIM content rendered by software developed by the US>Earth System Curator project integrated into the Earth System Grid Gateway (from http://earthsystemgrid.org, on the 17 April 2012). Elements of Simulation, Software, and Grid documents are shown. The box on the left shows some of the software component structure in a "tree-control", the title and abstract are those of the Simulation, and the Grid tab exposed is showing part of one of the grids used. A user can navigate around this representation of the CIM content to find out details of component properties, inputs and outputs etc.

GMDD
5, 1669–1689, 2012

Describing Earth System Simulations with the Metafor CIM

B. N. Lawrence et al.

```
<gridTile discretizationType="logically_rectangular">
     <description>Horizontal properties: The N96 Grid represents the 192-column horizontal
       coordinate system utilised within the Met Office Hadley Centre HadGAM1 and HadGAM2
       atmosphere models. This grid defines the horizontal locations of the physics (P) variables
       computed by these atmosphere models. The locations of U variables are offset by one-half of
       a grid cell to the east of P grid locations. The locations of V variables are offset by
       one-half of a grid cell to the north of P grid locations. Vertical properties: Vertical
       levels are terrain-following for levels up to k=29 and constant thickness above that
       level.</description>
     <extent>
       <latMin>-90</latMin>
       <latMax>90</latMax>
       <lonMin>0</lonMin>
       <lonMax>360</lonMax>
     </extent>
     <horizontalResolution description="1.875␣degrees␣in␣longitude␣by␣1.25␣degrees␣in␣latitude">
       <property>
         <value>145</value>
         <name>NumberOfLatitudinalGridCells</name>
       </property>
       <property>
         <value>192</value>
         <name>NumberOfLongitudinalGridCells</name>
       </property>
     </horizontalResolution>
     ...
</gridTile>
```

Fig. 4. The primary representation of CIM content is stored and exchanged in XML. This snippet of grid XML underlies some of the material shown in Fig. 3.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

GMDD

5, 1669–1689, 2012

Describing Earth System Simulations with the Metafor CIM

B. N. Lawrence et al.