



## **A 'Common Information Model' for the climate modelling process**

Allyn Treshansky, Gerard Devine, and the METAFOR Team

Department of Meteorology, Reading, United Kingdom (g.m.devine@reading.ac.uk)

The Common Information Model (CIM), developed by the EU-funded METAFOR project (<http://metaforclimate.eu>), is a formal model of the climate modeling process. It provides a rich structured description of not only climate data but also the “provenance” of that data: the software models and tools used to generate that data, the simulations those models implement, the experiments those simulations conform to, etc.. This formal metadata model is expected to add value to those datasets by firstly codifying what is currently found only in the heads of climate experts (the aforementioned provenance of climate datasets), and secondly by allowing tools to be developed that make searching for and analysing climate datasets a much more intuitive process than it has been in the past.

This paper will describe the structure of the CIM, concentrating on how it works with and what it adds to other metadata standards.

As alluded to above, current metadata standards concentrate on the contents of a climate dataset. Scientific detail and relevance of the model components that generated that data as well as the context for why it was run are missing. The CIM addresses this gap. However, it does not aim to replace existing standards. Rather, wherever possible it re-uses them. It also attempts to standardise our understanding of climate modeling at a very high level, at a conceptual level. This results in a UML description of climate modeling, the CONCIM. METAFOR extracts from this high-level UML the bits of the CIM that we want to use in our applications; These bits get converted into a set of XSD application schemas, the APPCIM. Other user groups may derive a different APPCIM (in a different format) that suits them from the same CONCIM. Thus there is a common understanding of the concepts used in climate modeling even if the implementation differs.

In certain key places the CIM describes a general structure over which a specific Controlled Vocabulary (CV) can be applied. For example, the CIM says things like "a component can have a child component" but it doesn't say anything about what types of components those can be. It is up to external CV to provide the content that users can use to fill in CIM instances. So there is a CV developed for CMIP5 which says things like "an atmosphere component can have a radiation component". This is a very powerful idea because it allows the CIM to support separate vocabularies for different user groups.

The CIM is divided into several different UML packages / XML schemas that work together. These packages describe several “first order” entities which can exist as stand-alone CIM documents: things like experiments, models, simulations, datasets, etc.. Internally, individual CIM documents can reference other documents. Datasets get used and generated by models which can be organised hierarchically and coupled together. Models are run in support of simulation activities. A simulation is performed within the framework of a particular experiment. That experiment will have certain prescribed requirements. The simulation must therefore conform to all those requirements.

This rich structure can be exploited by different tools being actively developed by METAFOR. Among these are a web-based tool that helps scientists to document their models and simulations in a CIM-compatible way, a search engine that allows users to submit a range of queries in order to find relevant (descriptions of) data and models and simulations, and a differencing tool that allows those descriptions to be compared against each other.

