The Earth System Curator: integration technology for models and data

Third Curator Workshop
Princeton NJ

V. Balaji (**balaji@princeton.edu**)[1]

[1] Princeton University

NOAA/GFDL

18 October 2007

Outline

1. What is a curator
   - Convergence of models and data
   - Model and grid metadata
2. Curator metadata architecture
3. Curator implementations
   - GFDL Curator and CDP Curator
   - FRE: a prototype for CRE
   - Links with ESMF and MAPL
   - Links to other projects
4. Summary
5. Issues for AR5

Outline

1. What is a curator
   - Convergence of models and data
   - Model and grid metadata
2. Curator metadata architecture
3. Curator implementations
   - GFDL Curator and CDP Curator
   - FRE: a prototype for CRE
   - Links with ESMF and MAPL
   - Links to other projects
4. Summary
5. Issues for AR5

The **routine** use of Earth System models in research and operations

Let's declare that 2000-2010 (the "noughties") is the decade of the coming-of-age of Earth system models.

Operational forecasting model-based *seasonal* forecasts delivered to the public;

Decision support models routinely run for climate policy, energy strategy, risk pricing.

Fundamental research the use of models to develop a predictive understanding of the earth system and to provide a sound underpinning for all applications above.

This will require a radical shift in the way we do modeling: an infrastructure for moving the building, running and analysis of models and model output data from the "heroic" mode to the routine mode.

Curators: a noughties technology

- The *comparative study of climate simulations* (e.g IPCC) across many models has spawned efforts to build uniform access to output datasets from major climate models, as well as modeling frameworks that will promote uniform access to the models themselves.
- The key element in the integration will be a curator. A curator begins begins with a crucial insight: that the descriptors used for comprehensively specifying a model configuration are needed for a scientifically useful description of the model output data as well. Thus the same attributes may be used to specify a model as well as the model output dataset: thus leading to a *convergence of models and data*.
- ESC – the Earth System Curator – is a pilot project building prototype elements of such a system. The current project is funded by NSF and brings together NCAR, Princeton, MIT and GA Tech.

Curators: a noughties technology

- The *comparative study of climate simulations* (e.g IPCC) across many models has spawned efforts to build uniform access to output datasets from major climate models, as well as modeling frameworks that will promote uniform access to the models themselves.
- The key element in the integration will be a curator. A curator begins begins with a crucial insight: that the descriptors used for comprehensively specifying a model configuration are needed for a scientifically useful description of the model output data as well. Thus the same attributes may be used to specify a model as well as the model output dataset: thus leading to a *convergence of models and data*.
- ESC – the Earth System Curator – is a pilot project building prototype elements of such a system. The current project is funded by NSF and brings together NCAR, Princeton, MIT and GA Tech.

Curators: a noughties technology

- The *comparative study of climate simulations* (e.g IPCC) across many models has spawned efforts to build uniform access to output datasets from major climate models, as well as modeling frameworks that will promote uniform access to the models themselves.
- The key element in the integration will be a curator. A curator begins begins with a crucial insight: that the descriptors used for comprehensively specifying a model configuration are needed for a scientifically useful description of the model output data as well. Thus the same attributes may be used to specify a model as well as the model output dataset: thus leading to a *convergence of models and data*.
- ESC – the Earth System Curator – is a pilot project building prototype elements of such a system. The current project is funded by NSF and brings together NCAR, Princeton, MIT and GA Tech.

Linking model and data frameworks

Community data frameworks (e.g ESG, an ESC partner) are under development, at various institutions, informally linked by the GO-ESSP. For model output data to be scientifically useful, the researcher must have some knowledge of how the data was produced. Model data requires a *model's eye view* description of the data, another layer of metadata, which might include:

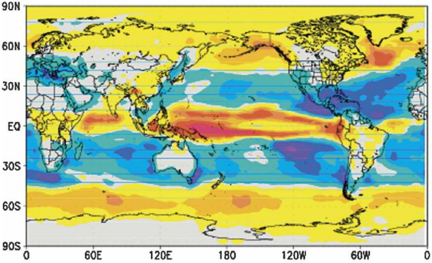
- Description of model components: e.g GEOS-5 atmosphere, land and sea ice coupled to MIT ocean.
- Description of grid configurations and resolutions.
- Choice of physics packages and input parameters.
- Model state and its fields.

ESMF and PRISM are emerging standards that allow the development of the model metadata layer, based on the state data structures and its base classes. (Think **`State`**, **`Grid`**, **`LocStream`**, . . .)

Semantic vs. syntactic, discovery vs. use

Descriptive metadata can be succinct, and can be used to discover certain aspects of the data. But almost any serious use requires deeper knowledge. The boundary between *discovery* and *use*, *semantic* and *syntactic*, is blurred by the use of controlled vocabularies and ontologies.

Graphics such as this from Held and Soden (2006) are so routinely produced from the IPCC AR4 database that we've ceased to marvel at it. This is a composite of output from 20 models worldwide, run with minimal coordination.



Model and grid metadata

Physical fields: standard vocabulary for describing the relevant physical quantities (viz. CF **`standard_name`**).

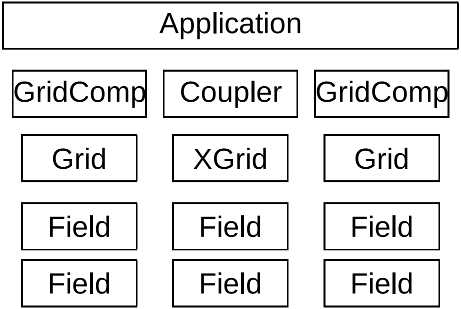
Geospatial information: location information: latitude, longitude, elevation. This set of standards unites a much larger community (mobile phones, GIS), in which our community has begun to play a role. We can provide some useful extensions toward 3D and 4D data.

Grid structure: interrelations between grids, between points and grids. With this information available, it is perhaps possible to perform regridding and subsampling of data by user request, on the archive servers.

Model metadata: describing data source comprehensively, relatively easy for observations, harder for models but can asymptote toward completeness starting from current PCMDI standard. Two levels of model metadata: components and applications.

Model metadata

- Application metadata: experiment, scenario, institution, contact: currently covered by CF/CMOR.
- Component metadata: physical description of component: currently covered by CMOR, extended by NMM.
- Coupler metadata: inventory of export and import fields, interpolation methods. Currently covered by OASIS4 XML, not exported to model output. Associated with an XGrid.



Application

GridComp Coupler GridComp

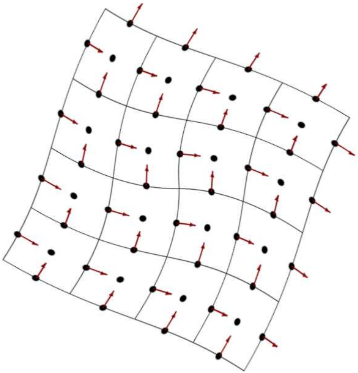
Grid XGrid Grid

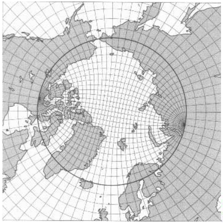
Field Field Field

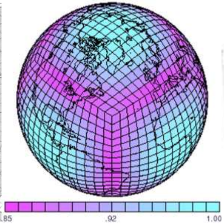
Field Field Field

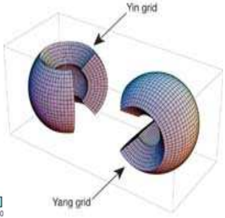
Grid metadata

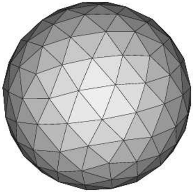
The Mosaic Gridspec is now under consideration as a draft CF standard. Contains information for differentiation, integration and regridding on generalized grids. Currently implemented for coupling models at GFDL (cubed-sphere atmosphere, tripolar ocean, lat-lon land and river grids) and as an XML schema for data web services in the European Genie project.









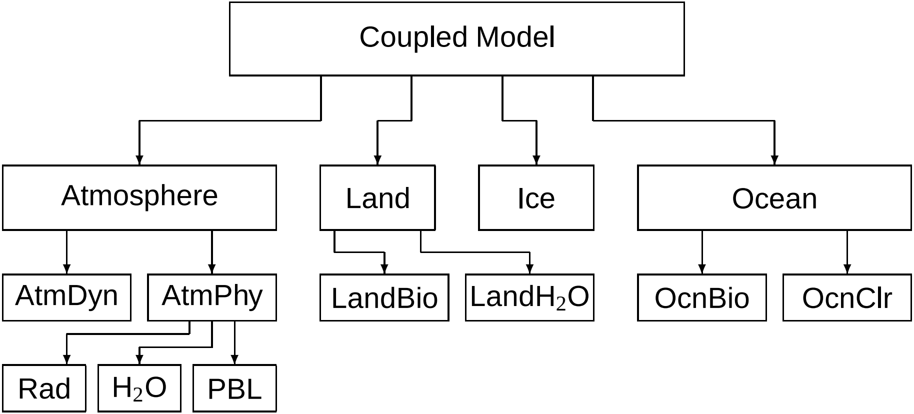


Outline

1. What is a curator
    - Convergence of models and data
    - Model and grid metadata
2. Curator metadata architecture
3. Curator implementations
    - GFDL Curator and CDP Curator
    - FRE: a prototype for CRE
    - Links with ESMF and MAPL
    - Links to other projects
4. Summary
5. Issues for AR5

Model architecture

The construction of complex Earth system models out of *components* is now commonplace in the design of modeling software. ESMF (US) and PRISM (EU) are emerging standards for making interoperable model components.



Component specification

A component is specified using the following schemata:

- **NMMComponent** scientific, technical, and numerical (computational) properties;
- **CIAO** component input and output: potential coupling fields, can it also be used for diagnostic output?
- **gridspec** comprehensive grid specification;
- **parameter, parameterGroup, inputDataset** : scientific configuration;
- **platform** technical configuration.

Model specification

A model is assembled out of available components dispensed by a curator. The component specification can be used for *compatibility checking* before assembly.

PotentialModel a list of NMMComponents.

- SMIOC specification of coupling fields (and output fields)?
- gridspec mosaic of component gridspecs;
- framework software infrastructure (e.g ESMF). A framework may be associated with a runtime environment.

Application specification

An application is a coordinated campaign of experiments run by many models

NMM Application : Describes a grouping of models/experiments/simulations; It is envisioned that the scientist would create a NMM Application XML file to "bin" or "wrap" a set of models into one application.

ESG Ontology described a set of resources whereby the outcome of an experiment can be perused.

ESC target schema

The ESC target schema embodies the elements described above in the ESC metadata architecture.

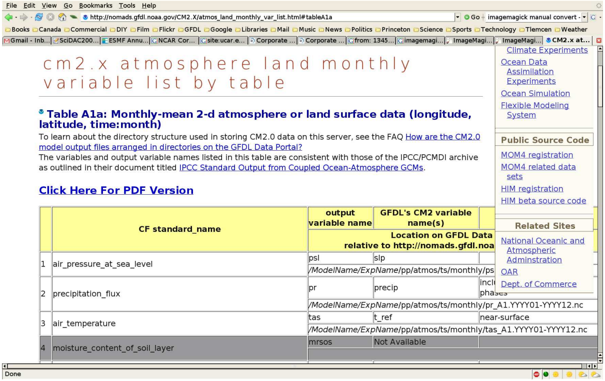
`http://earthsystemcurator.org/index.php?option=com_content&task=view&id=53&Itemid=85`

The complete schema, including all the use metadata required by the runtime environment, is quite huge (see FRE schema and underlying RDB in tomorrow's talk). We propose to define various use profiles or subsets for various applications, e.g compatibility checking, AR5 model metadata, FRE.

Outline

1. What is a curator
   - Convergence of models and data
   - Model and grid metadata
2. Curator metadata architecture
3. Curator implementations
   - GFDL Curator and CDP Curator
   - FRE: a prototype for CRE
   - Links with ESMF and MAPL
   - Links to other projects
4. Summary
5. Issues for AR5

GFDL Curator and CDP Curator



File Edit View Go Bookmarks Tools Help
http://nomads.gfdl.noaa.gov/CM2.X/atmos_land_monthly_var_list.html#tableA1a
Go
imagemagick manual convert
Books Canada Commercial DIY Film Flickr GFDL Google Libraries Mail Music News Politics Princeton Science Sports Technology Tlemcen Weather
Gmail - Inb… SciDAC200… ESMF Annu… NCAR Cor… site:ucar.e… Corporate … Corporate … from: 1345… imagemagi… ImageMagi… ImageMagi… CM2.x at…
Done

cm2.x atmosphere land monthly variable list by table

Climate Experiments
Ocean Data Assimilation Experiments
Ocean Simulation
Flexible Modeling System

Public Source Code

MOM4 registration
MOM4 related data sets
HIM registration
HIM beta source code

Related Sites

National Oceanic and Atmospheric Adminstration
OAR
Dept. of Commerce

Table A1a: Monthly-mean 2-d atmosphere or land surface data (longitude, latitude, time:month)

To learn about the directory structure used in storing CM2.0 data on this server, see the FAQ How are the CM2.0 model output files arranged in directories on the GFDL Data Portal?

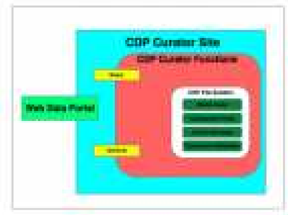
The variables and output variable names listed in this table are consistent with those of the IPCC/PCMDI archive as outlined in their document titled IPCC Standard Output from Coupled Ocean-Atmosphere GCMs.

Click Here For PDF Version

	CF standard_name	output variable name	GFDL's CM2 variable name(s)	
		Location on GFDL Data relative to http://nomads.gfdl.noa		
1	air_pressure_at_sea_level	psl	slp	
		/ModelName/ExpName/pp/atmos/ts/monthly/ps		
2	precipitation_flux	pr	precip	incl phases
		/ModelName/ExpName/pp/atmos/ts/monthly/pr_A1.YYYY01-YYYY12.nc		
3	air_temperature	tas	t_ref	near-surface
		/ModelName/ExpName/pp/atmos/ts/monthly/tas_A1.YYYY01-YYYY12.nc		
4	moisture_content_of_soil_layer	mrsos	Not Available	

GFDL Curator currently creates dynamic data catalogues from metadata (currently not from the Curator schema itself, but from metadata embedded in the datasets: integration with schema is underway).

The CDP Curator provides an interface both query and archival of ESMF components.



Operational use of model frameworks

The next stage in the evolution of frameworks is the addition of a *runtime environment*.

- Source code maintenance across many repositories;
- Model configuration, launching and regression testing encapsulated in XML descriptors;
- Relational database for archived model results;
- Standard and custom diagnostic suites;
- Branching: "descent with modification".

FRE: a prototype for CRE

The FMS Runtime Environment (FRE) describes all the steps for configuring and running a model jobstream; archiving, postprocessing and analysis of model results.

`fremake, frerun, frepp, frecheck, ...`

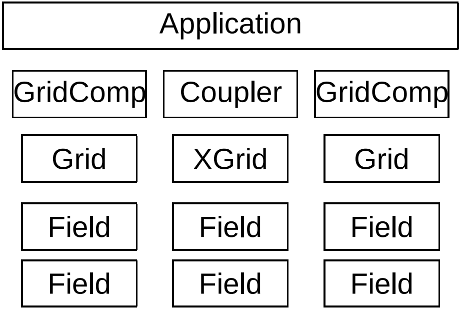
The Regression Test Suite (RTS) is a set of tests that are run continuously on a set of FMS models to maintain and verify code integrity.

FRE was successfully used at GFDL for the development of climate models targeted for IPCC (CM2.0 and CM2.1) and management of GFDL's IPCC data. We are currently merging FRE and ESC schema to make FRE serve as a prototype Curator Runtime Environment.

`http://www.gfdl.noaa.gov/˜fms/fre`

ESC links to ESMF and MAPL

- ESMF data structures (**`Grid`**, **`Field`**, **`LocStream`**, **`ConfigAttr`**, **`State`**, . . .) are ideal containers for holding the metadata. Tools are under development for extracting the Curator metadata from ESMF components registered under the Community Data Portal.
- MAPL specifications of coupling will be encoded in Curator coupler metadata, which itself is being developed on the basis of the OASIS4 (PRISM) schema.



Application

GridComp | Coupler | GridComp

Grid | XGrid | Grid

Field | Field | Field

Field | Field | Field

ESC links with other projects

- Earth System Grid is developing key technology for serving data from multiple modeling centers (IPCC, NARCCAP), and are partners in the metadata definition effort focused on AR5.
- Numerical Model Metadata (NMM) based in the University of Reading defines discovery metadata for Earth System Models, and have converged with ESC schema where there is overlap. The Metafor proposal seeks to formalize the relationship further.
- The FLUME project at the UK Met Office is developing similar metadata and is interested in looking at auto-generation of FLUME "glue" using the common information model. Also part of Metafor.

Outline

1. What is a curator
   - Convergence of models and data
   - Model and grid metadata
2. Curator metadata architecture
3. Curator implementations
   - GFDL Curator and CDP Curator
   - FRE: a prototype for CRE
   - Links with ESMF and MAPL
   - Links to other projects
4. Summary
5. Issues for AR5

Summary

- Curators are a natural outgrowth of frameworks. By blurring the edges between models and data, between objects and services, they are a typical "fuzzy boundary" technology. (Roger Sessions, *ACM Queue*, 2004: *Fuzzy Boundaries: Objects, Components, and Web Services*.)
- The ESC project is drawing up a metadata architecture which can aggregate within a single information model, metadata layers (e.g component, coupler, campaign) to be assigned to different domains of expertise.
- A functional schema-driven runtime environment allowing composition, configuration, running, archival, and analysis of Earth system model data is already a (limited) reality, and we are currently probing its limits and imposing generalizations.

`http://www.earthsystemcurator.org`

Outline

1. What is a curator
   - Convergence of models and data
   - Model and grid metadata
2. Curator metadata architecture
3. Curator implementations
   - GFDL Curator and CDP Curator
   - FRE: a prototype for CRE
   - Links with ESMF and MAPL
   - Links to other projects
4. Summary
5. Issues for AR5

Issues for AR5

- Native grid data (Curator/Metafor for spec; originating site takes responsibility for regridding algorithm; who deploys it as a web service?);
- Increased use of forcing fields and initial condition fields (Curator/Metafor for spec; CMIP4 for content?);
- Are the actual stored “naked” files going to be useless without metadata or data transformations?
- Is the system for exchanging metadata going to be ready in time?
- Is the system for exchanging metadata going to be fault-tolerant? Who is responsible for failures to hand off?
- Who is responsible for the software stack at the server node?