

NSF 04-528 Information Integration

Earth System Curator: Spanning the Gap Between Models and Datasets

PRINCIPAL INVESTIGATOR

Cecelia DeLuca NSF National Center for Atmospheric Research
Scientific Computing Division
cdeluca@ucar.edu
303-497-1828

CO-INVESTIGATORS

V. Balaji NOAA Geophysical Fluid Dynamics Laboratory/Princeton University
Modeling Systems Group
v.balaji@noaa.gov
609-452-6516

Don Middleton NSF National Center for Atmospheric Research
Scientific Computing Division
don@ucar.edu
303-354-1250

Spencer Rugaber Georgia Institute of Technology
spencer@cc.gatech.edu
404-894-8450

John Marshall Massachusetts Institute of Technology
Department of Earth, Atmospheric and Planetary Sciences
marshall@gulf.mit.edu
617-253-9615

1 Project Summary

Climate research has been given urgency and direction by the imminent social, economic and political implications of observed climate change [e.g. 1,2]. Complex, multi-component models are being used at many national centers and universities to develop future projections of climate, and a key goal of current research is to reduce the uncertainty of these projections by understanding differences in model output. This *comparative study of climate simulations* has spawned efforts to provide uniform access to output datasets from climate models, and to build modeling frameworks that facilitate component exchanges and comparisons.¹ However, models and datasets are currently treated within the climate community as distinct and separate entities, an artificial barrier that inhibits access to resources and results. We propose to unify the treatment of models and datasets by developing a common language - a metadata formalism - with which to describe the two, and by prototyping a set of tools based on that formalism that allows the researcher to manipulate models and datasets seamlessly and with ease. The goal, in the end, is to increase the productivity of climate researchers and our understanding of the Earth system.

The work proposed builds on two ongoing community efforts, the Earth System Modeling Framework (ESMF) [3,4] and the Earth System Grid II (ESG) [5]. The ESMF is a national initiative to develop common modeling infrastructure for the nation's climate and weather models. It involves scientists, computer scientists, and software engineers from NSF, DoD, NASA, NOAA, and DOE, as well as researchers and students from MIT, UCLA, and the University of Michigan. It will be the technical foundation for the emergent NASA Modeling, Analysis, and Prediction (MAP) Climate Variability and Change program and the DoD Battlespace Environments Institute (BEI), as well as a host of climate and weather models. In order to manage information about the models and components adopting ESMF, the project team has developed an on-line Component Database whose structure reflects the hierarchical, component-based composition of models using the framework [6].

The primary goal of ESG is to make the output of high-resolution, long-duration simulations performed with advanced DOE SciDAC/NCAR climate models available to global change impacts researchers nationwide. ESG is developing and integrating a range of Grid,² data/metadata, and portal technologies, including the OPeNDAP [7] remote access protocols, Globus Toolkit [8] technologies for authentication, resource discovery, and resource access, and Data Grid technologies developed in other projects. It involves computer scientists at five laboratories (ANL, LBNL, LLNL, NCAR, ORNL) and one university (ISI), working in close collaboration with climate scientists at several sites.

The proposing team seeks to explore those aspects of ESMF and ESG that can be usefully aligned, and to prototype a new entity, the Earth System Curator, that spans the gap between the two. The Curator 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. The *convergence of models and data*, and the development of a common metadata schema that describes both, will be the basis for this unique and powerful community resource. The Curator will provide a community database from which researchers can archive and query a wide class of Earth system models, experiments, model components, and model output data and results. Researchers will subsequently be able to either analyze model output from pre-existing runs, or access a model and modify and run it themselves, either on a local computer or on the virtualized resources of the computational Grid. In addition to the query function, we will prototype a tool that will test if sets of model components or datasets can viably interact to form an application. We intend to handle technical compatibility of model components first, and later explore bounds and techniques for expressing and addressing scientific compatibility. Finally, as part of the Curator effort we will prototype tools for auto-generation of component wrappers and applications.

¹ A researcher may, for example, want to experiment with different ocean models coupled to the same atmosphere.

² We will use Grid to refer to virtualized computational resources, and grid to refer to model grids.

2 Current Funding and Related Project Objectives

ESMF began in 2001 as a NASA-funded project, and has since transitioned to multi-agency funding, under the coordination of the ESMF Interagency Working Group. ESG, which also began in 2001, is sponsored by the DOE Scientific Discovery Through Advanced Computing (SciDAC program). Both ESG and ESMF have clear missions within the climate community, and their current funding focuses on the development of production-quality systems. Neither project currently has resources dedicated to exploring research issues that may lead to next-generation systems, or to examining and forging vital links between these projects. Funding for the Curator effort will help to bridge the gap, both between the ESG and ESMF projects, and between the current instantiation of these projects and the modeling environments of the future.

3 Technical Merit

The proposed work will benefit both Climate Science and Information Technology (IT). With regards to Climate Science, the Earth System Curator will provide a uniform view to users of both climate models and model data. In practical terms, it will include a query tool enabling access to models, experiment configuration details, and output data; a compatibility tool to test the feasibility of component and dataset combinations; and an auto-generation tool for application assembly. When combined with ESMF and ESG, the Curator offers the first steps towards an end-to-end, fully integrated problem solving environment that has the power to transform the way climate research is performed, day-to-day. The transformation will be in the ability to query a wider range of information sources, to target searches and find information rapidly, to use that information together with compatibility and auto-generation tools to configure and run new experiments, locally or virtually, to store and disseminate information about results, and to analyze and visualize results to acquire new knowledge (and formulate new questions).

With regards to IT, the Curator is an example of the integration of large, recently established modeling and data infrastructure efforts, at a scale that is challenging both technically and socially. We will examine how the metadata used for applications (in the ESMF Component Database, and CCSM [9], MIT [10], and GFDL [11] climate models) and datasets (in ESG) differ, how they can be unified, and how a hierarchical component paradigm can be expressed as metadata. We will explore the implementation, interaction and combination of software component technology, database technology, and code auto-generation for a real-world problem. The Curator effort will also explore the implications of developing abstractions in a computational environment where performance considerations dominate.

4 Broader Impact

Defining the future of IT in Climate Science The Curator is part of a community vision for the use of IT in climate and related research that has been advanced by a large collaboration that includes leading modelers from five agencies and a number of universities. This vision is articulated in a collaborative white paper, which introduces the Curator as the basis for an advanced Earth System Modeling Environment (ESME) [12]. The Curator prototype will help to further suggest and define the form of next-generation modeling and data management tools, by offering a concrete representation to add credibility to innovative ideas. Further, the ESMF and ESG Co-Investigators, by virtue of their projects' emphases on production software and extensive customer bases, are in an excellent position to transition the Curator tools into a viable product following an NSF-funded prototype stage. Over the course of the Curator effort, many researchers associated with ESMF and ESG will be encouraged to try out and offer feedback on the Curator software.

IPCC and MIPs The Intergovernmental Panel on Climate Change (IPCC) [13] as well as many other Model Intercomparison Projects (MIPs) [e.g., 14,15] describe standard experiments that are independently run by diverse modeling groups. The process wherein each group attempts to set up an identical experiment is currently laborious. As part of its auto-generation research, the Curator will prototype a mechanism that can specify an experiment in XML, with no reference to an individual model. A researcher required to conform to a particular model configuration for a MIP will be able to use this tool to specify and configure their model according to the given constraints. Another key feature we intend to prototype, as part of the Curator database and query tool, is the archival of publicly available MIP models in a form that allows modeling groups to easily and independently access and manipulate them. The model data analogue to this capability is already within the scope of the ESG; we shall investigate the capability to access models the same way.

Related domains People in related science domains often look to the developers of climate models and modeling infrastructure as leaders in community organization and the use of high performance computing (HPC). Because climate modelers require the contributions of disparate groups and enormous computational resources for even basic applications, they were pioneers of parallel, multi-component systems and community models. Fusion, sedimentation, space weather, and solid Earth groups are among those who have approached Curator-related projects such as ESMF and the GFDL Flexible Modeling System (FMS) [11] seeking technical and organizational advice. Advances achieved with the Curator project will influence other domains through conferences and publications, and through a web of relationships founded on a shared need for multi-component HPC modeling, ease of information archival and access, and similarities in simulation numerics.

Policy and social impacts Enhancements in scientific productivity that lead to improved predictive capability can also lead to better information for policy-makers, and the ability to formulate appropriate responses to current and potential climate changes. While we expect the advances in climate prediction due directly to the Curator effort to be both difficult to track and modest at best, we do feel that an integrated environment for Earth system research is critical to addressing world climate issues in the near future, and the Curator is a definitive step in that direction.

Engineering students and the general public At Georgia Tech, the College of Computing offers a variety of Software Engineering (SE) courses, both undergraduate and graduate. A key feature of these courses is their strong project orientation. That is, lecture material is reified as student activities during actual project development. Historically, these projects have focused on user interfaces that collect information and requests and interact with data bases. The reason for this is that such projects require relatively little domain knowledge to be learned by the students. However, Georgia Tech is primarily an engineering institution, and it would be desirable to incorporate more engineering-based projects into the SE curriculum, if only the domain-knowledge problem could be overcome. The Curator provides one way to do this, by encapsulating access to a wide variety of Earth system models and data, and by providing a sophisticated interface by which new applications can be incorporated. As part of the proposed effort, the Curator will be incorporated into the project infrastructure for SE courses. Appropriate introductory material will be prepared and project opportunities defined. Many of the project opportunities will take the form of making climate data accessible to the general public. In this way the Curator not only provides SE students an opportunity to participate in an actual ongoing engineering development effort, but the resultant projects can explore making Earth science more available to the general public. Experience with the approach, along with curriculum materials, will be made available to other educational institutions via web publication and presentation at SE education conferences.

Observational Data Community NASA and NOAA data assimilation projects have participated actively in the development of ESMF, are represented in the project management structure and are framework customers. The ties between these groups, the observational data community, and the Curator project

will be maintained and grown so that there is active communication between Curator project members working on metadata formalism, and existing and planned work on the part of the observational community on metadata for observational datasets.

Publication Scientific publication of model results has the potential to be transformed by the Curator. One can envision a future where model datasets will be annotated with references to published results; conversely, publications will carry references to where models and data are to be accessed, so that a reader could in principle carry out independent analyses, or even launch new model runs to extend the published results.

5 Project Description

5.1 A Curator Use Case

Consider the following scenario. A climate impacts scientist has developed a malaria model where mosquito breeding rates can be modeled as a function of local temperature and rainfall in swamp ecosystems. She wishes to use her model to understand how malaria pandemics might unfold in a warming climate. Since there is no climate feedback from the mosquitoes to the climate, her model could be run “offline” from a climate dataset. She finds and downloads a suitable model dataset from among models worldwide that have run “IPCC 2006 climate projections” and runs her experiment. In the course of her study she discovers that other scientists, using the same data, have shown that tropical rainfall is systematically under-predicted by the model she is using. They have suggested how to remove the bias, but their data are not archived. Though not a climate simulation expert, she is able from the available information using the Earth System Curator to reconfigure the model and rerun it with the rainfall bias removed and generate new data for her malaria study.

This scenario is currently fiction, but is not as far from being fact as one might suppose. Some elements of what is needed to get there are already in place, but a critical piece of the cyberinfrastructure puzzle remains to be designed and built. This proposal aims to develop a prototype of the information integration system that is missing.

5.2 Community Divisions in Earth System Research

Models and model output data sets represent essentially the same information: a model code, its boundary values, its physical inputs, and its configuration parameters can be viewed as a compressed form of the data set generated when the model is executed. When addressing a scientific question, a scientist may find it useful to retrieve a previously generated data set, the model that generated that data set, observational data describing the time and region of interest, or any combination of these.

Despite their affinities, models and data sets tend to be treated as separate entities in the climate and weather domain. The model data infrastructure (MDI) community, whose raw material is model data that has already been generated, develops tools for archiving, describing, searching and retrieving that data. The MDI community is represented by efforts such as the NetCDF Climate and Forecast (CF) Conventions [16] and the Earth System Grid II (ESG) [5]. These groups have worked for a number of years to establish descriptors for Earth system model data sets in the form of conventions and standards. Newer work focuses on ontologies, schema that encompass both traditional metadata and relationships between descriptors [e.g., 17]. The overriding goal for these groups is to allow researchers to archive, locate and mine data sets of interest. Figure 1 depicts a screen shot of the ESG web portal interface to climate simulation data.

The observational data infrastructure (ODI) community, represented by efforts such as HDF-EOS [18], faces similar issues. The metadata required to describe observational data is further complicated by factors including the irregularity of data locations, the possibility that these locations may change over time (e.g. satellite swaths, ocean floaters), and instrument characteristics, limitations, and biases that must eventually be accounted for. Like the MDI community, the goal of the ODI groups is to allow for the efficient archival, description, and retrieval of data.

The modeling infrastructure (MI) community, represented by projects such as the Earth System Modeling Framework (ESMF) [3], the GFDL Flexible Modeling System (FMS) [11] and the Common Component Architecture (CCA) [19], approaches the problem differently. The Earth system consists of a complex hierarchy of physical entities – atmosphere, ocean, and so on – with complex feedbacks at multiple space and time scales. These physical entities themselves contain processes that can be independently modeled: atmospheric radiation, cloud processes, ocean biogeochemistry, and so on. As the physical representations increase in number and complexity, model designers look to the idea of a model *component* as a software abstraction. The ESMF and related projects follow similar hierarchical component-based design patterns to describe Earth system model components in standard ways. Component-based architectures allow researchers to create generic components based on interface rules and slot them interchangeably into an application. This approach ameliorates the technical – though not the scientific – difficulties involved in coupling multi-component systems. As an example of a component hierarchy see Figure 2, which shows the structure of the new Goddard Earth Observing System model, version 5 (GEOS-5). GEOS-5 has been developed using ESMF from the ground up. It will be used for research in satellite data utilization; as part of an atmospheric data assimilation system; for weather, sub-seasonal, and seasonal to interannual forecasting; for atmospheric chemistry studies; carbon cycle research; and research on ocean-atmosphere and atmosphere-land interactions.

A feature that many modeling frameworks share, and that is a key to the proposed work on Curator, is the ability to store metadata – often, descriptors of physical fields - within framework data structures such as components. Thus, when two components seek to share a boundary field called “Temperature”, the descriptive information makes it possible to check if both components use the same units for temperature. Model output is another function performed by the modeling framework: this layer of software accesses the stored metadata and writes it as output.

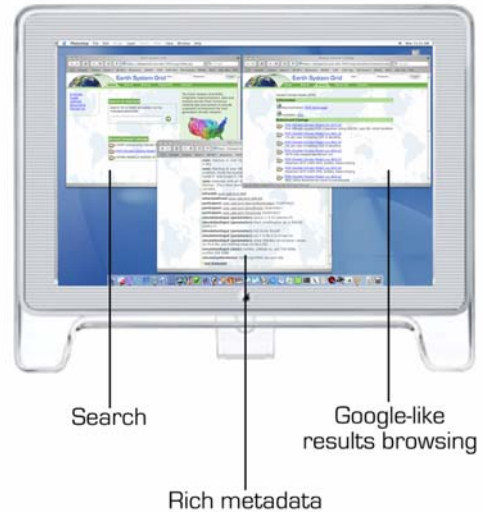


Figure 1 *ESG Portal Interface*

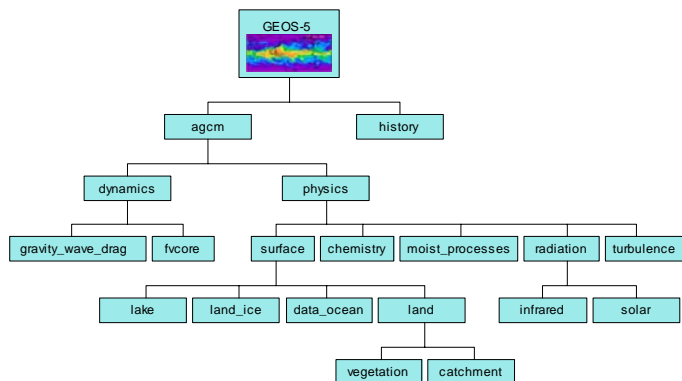


Figure 2 ESMF-based architecture of the GEOS-5 atmospheric general circulation model. Each box is an ESMF component. Component and coupling interfaces are standardized to facilitate exchanges and extensions. The operations in each component or coupling transformation can be easily customized.

5.3 The Case for Integration of Models and Datasets

Just as it is necessary to describe and archive data sets in order to understand and use them, so it is necessary to describe and archive modeling components in order to understand and use them.

Although it is not a primary function of the ESMF project to archive information about components and applications, the difficulty of tracking the mix of codes involved in the ESMF collaboration motivated the project team to develop an ESMF Component Database [6]. This stores both information about the individual fields that a particular component or application requires and can make available, and information about the component or application as a whole, such as the sorts of calendars it supports and what platforms it can run on. The ESMF Component Database is hierarchical, so that, for example, the fields that can be exported by a parent component in the hierarchy potentially include all those fields that can be exported by its children.

When developing metadata for modeling components, as we have done in the ESMF Component Database, it immediately becomes clear that there are strong similarities between this task and the task of crafting metadata for the description of model data sets. We see the analogy, for example, in the need to describe individual physical fields. Components must be able to communicate to other components descriptions of the fields they require and the fields they can export while the application is executing. A model dataset must describe the fields that the set of components that comprise the modeling application can export – essentially the same field information, destined for an end-user rather than another component.

It is to the advantage of scientists, who rely upon changing and improving models in order to make scientific advances, to be able to reference and retrieve data sets and the modeling components and configurations used to generate them in a seamless fashion. It is also advantageous for the MI community to leverage the progress made by the MDI and ODI communities, both in the area of metadata and approaches to information archival and retrieval. Finally, it makes sense for these communities to respond to upcoming challenges in a coordinated fashion. These challenges include utilization of the computational Grid, an area which has obvious benefits for the MDI community and one in which it has taken the lead; and instrumentation of models so that metadata can be generated consistently and automatically.

We propose to take the first steps towards developing a unified approach to metadata description and auto-generation for modeling components and model data sets. We also plan to engage key members of the ODI community in order to examine how the issues that we encounter relate to their efforts.

5.4 Research and Development Activities

The key activities in this proposal are summarized as follows:

1. Based upon an in-depth understanding of how the metadata-laden data structures of ESMF are used by three key Earth system model applications (CCSM, FMS, MITgcm), a comparative study of metadata efforts within the broader modeling community, and the initial implementation of the ESMF Component Database, outline a formal metadata schema to describe Earth system models and model components;
2. Understand and define the ontology relating these ESMF-based model schema and the ESG schema for model data;
3. Build a prototype relational database (Earth System Curator) based upon this ontology, and a corresponding set of query tools;
4. Prototype a compatibility tool for determining whether a selection of components and datasets can form a viable application, focusing first on straightforward technical aspects such as platforms supported and later exploring methods for expressing and reporting scientific compatibility; and
5. Investigate approaches to auto-generating wrappers for components and couplers, as well as applications that consist of assemblages of components and datasets, using the Curator database and compatibility tool. One of the features desired is the ability to specify an application template and later instantiate it with a particular set of components. While it is not a primary focus of this project, we would also like to explore how assembled ESMF-compliant applications could be launched virtually using Curator by leveraging the underlying Grid capability of ESG. The goal is to develop an integrated end-to-end capability for scientific investigation.

These steps are described in greater detail in the following sections. Figure 3 illustrates the elements of the full Curator system.

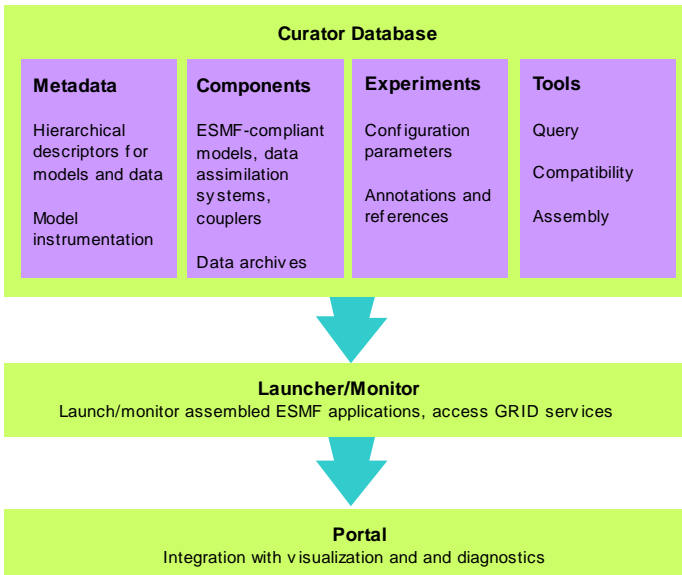


Figure 3

Curator schematic showing the elements of the full system: metadata formalism, ESMF-compliant components, stored experiments, and a variety of tools. We plan to base the application launcher and the initial portal user interface on existing capabilities in ESG.

5.4.1 Definition of Model Metadata

Earth system models can broadly be described as composed of components in which physical quantities are integrated on a physical grid. In a framework like ESMF, these are described in terms of successive layers of abstraction consisting of metadata-laden data structures. These layers are:

grid: the physical grid expressed in a standard way, so that component-neutral regridding software can be used to transform quantities from one grid component to another, with no knowledge of the components themselves. We seek, as part of this proposal, to inscribe the grid metadata within community standards and conventions such as CF, so that analysis tools cognizant of these conventions may take advantage of grid information.

field: a physical variable discretized on a `grid`, along with metadata describing the physical quantity itself. The field metadata in ESMF have been designed to resemble those within the CF convention, so that CF-compliant model output may be produced if desired.

state: the instantaneous state of some set of fields within a model component. Typically these are used as part of “import” and “export” states that are exchanged between components; but they are often used to contain the entire model state as well.

attribute: configuration attributes of a component: these are generic, but are intended to contain all the physical input parameters used to configure a model.

component: a top level computational entity. Components are hierarchical; that is, they may be composed of other components. The top-level component is the application or model itself.

These software layers exist in the ESMF, and ESMF-compliant applications in the near future will be using these abstractions, rich in metadata, to describe a wide range of models across the weather and climate community. Simply by using these abstractions and encoding them systematically in model output, we are creating a layer of formal, structured, hierarchical metadata. We call this the *model metadata layer*, and it is the core of the Curator. The model metadata layer is what makes possible for either a model configuration or a model dataset to be the result of a database query.

5.4.2 Comparative Study of Model Metadata

We propose to undertake a comparative study of model metadata from many institutions to understand the range of variation among components from diverse sources. Understanding component diversity is increasingly urgent. There is a controversy within the community about the feasibility and advisability of treating components as interchangeable bits of code that can be slotted together at will. An understanding of component diversity will mark the limits of such an approach. We seek to answer two questions:

- Are components ostensibly labeled “atmosphere,” say, sufficiently similar that a physical interface may be well-defined? Or, to put it another way, to what extent do two such components share a state?
- Do different modeling organizations see component granularity the same way? What incompatibilities are introduced if one model treats atmospheric chemistry say, as an indivisible entity within an atmosphere component, whereas another treats it as an independent component?

We propose that three of the modeling groups that are participants in ESMF (NCAR CCSM, PU/GFDL FMS, and MITgcm) become the core of the group defining model metadata. Although we do not have a CCSM Co-Investigator on this proposal, the CCSM project is actively engaged in metadata definition

activities in collaboration with GFDL and ESG, and we expect interactions with the Curator effort to be a natural and modest extension.

5.4.3 Definition of Technical Metadata

Modeling components have restrictions in terms of platform portability, permitted parallel programming models, and so on. The software abstraction within ESMF that describes computational resources is the virtual machine or vm. The vm is not a new concept, but is a useful one [e.g. 20, 21, 22].

We outline the function of this additional metadata layer as follows:

vm: the vm describes the range of current programming models: distributed, shared or hybrid memory, using shared-memory, message-passing, or remote-memory access semantics – and does so in a uniform way.

It is possible to use the vm technical metadata to ascertain which components may usefully be assembled and run on a given platform.

5.4.4 Model Metadata Acquisition and the Compatibility Tool

Based upon model metadata and technical metadata, we will create a component registry wherein model metadata is collected. There are two possible approaches:

- a *registry tool*, where the model developer manually enters the information that makes up the model metadata;
- or a *source-scan tool*, which uses knowledge of ESMF data structures to extract this information automatically from existing components.

ESMF components are well-structured enough to make both approaches feasible. The registry will be part of a compatibility tool that is likely to be a hybrid of both approaches, a machine pass followed by a human pass. The tool will determine whether:

- it is *technically feasible* to use a component in an application: does it run on the target platform, and so on?
- it is *physically feasible* to use a component: are the range of available resolutions sufficient for the problem at hand; do the physical subcomponents match the problem under study, etc.?
- it is *compatible* to use a component with other components in the application: do the available output fields match the required input field of the other components, can all components share a common calendar, etc?

5.4.5 Formal Basis for Discovery Metadata and the Query Tool

The ESG describes model output data in terms of two levels: **usage metadata** describing the data at the physical variable level; and **discovery metadata** higher-level information about the origins of the data: model, version, institution, type of simulation, etc. This is currently somewhat freeform. As part of this project, we seek to use the model metadata as a formal substrate for inscribing discovery metadata about model output data. Key tasks here include:

- Identify and correct incompatibilities between current usage metadata within the ESG community and the model metadata that results from use of the metadata-laden data structures in ESMF. A specific example of what will result is the formal adoption of a convention for describing the grid in an application, so that analysis tools automatically acquire comprehensive knowledge of the grid. Such tools are currently restricted to the simplest regular, aligned grids: the ESMF grid abstraction will eventually provide a means of extending this to much higher levels of complexity, including generalized curvilinear coordinates and unstructured grids.
- Design a prototype database that uses the model metadata and unites models, model components, and model data.
- Develop a prototype query tool, based on the Curator, capable of returning as appropriate references to either modeling components or model data sets.

5.4.6 Putting it All Together: The Auto-Generation Tool

The Curator team will explore mechanisms for the auto-generation of component wrappers, as outlined in Section 5.5.6. Components may be science modules, couplers, or datasets. This virtualization of components will be a useful and powerful capability, as many researchers like to be able to swap in both “live” and “data only” versions of particular modules in climate and related models. It is currently difficult to introduce such options. The Curator will offer a standard way of bringing scientific data into an application as an option instead of a “live” model – and with integrated access to the datasets archived in ESG, a tremendous store of dataset options.

Using components with automatically generated wrappers, hand-adapted components, the metadata formalism we develop, and the compatibility tool, we will look at ways to generate applications out of components that are found to be technically compatible. We will also examine how to create applications that fit a particular configuration template, modeling our investigation on the kinds of constraints imposed in climate Model Intercomparison Projects (MIPs) [e.g. 14]. We are interested in exploring how such applications can then be stored and annotated for distribution in the Curator database, which we hope will eventually become an extension of ESG. The underlying Grid technology of ESG suggests the next step, which is to use the Curator/ESG combination to run the assembled climate models in a virtualized manner. We will examine the possibilities for doing this, capturing our results in a prototype tool that will be presented to the scientists involved in ESMF and ESG for feedback on its implementation and its potential usefulness.

5.5 Information Technology Research Questions

The proposed work raises a variety of research questions in the IT area.

5.5.1 Metadata Framework

The Consultative Committee for Space Data Systems (CCSDS) has been a leader in coordinating the development of standards for archival of digital information. The Reference Model for an Open Archival Information System (OAIS) [____] has been approved as an ISO standard, and it has provided a framework for a number of efforts that are highly relevant to the research proposed here. Particularly relevant is the effort on *A Metadata Framework to Support the Preservation of Digital Objects* by The OCLC/RLG Working Group on Preservation Metadata [25]. This framework provides recommendations for storing data content and representation information, including content description, and software and hardware

5.5.2 Metadata Registry

The Curator project will need to manage the creation and evolution of the very standards by which metadata is produced, maintained, and used. It will not be sufficient to merely provide a repository into which metadata can be stored and subsequently retrieved. The Earth System Curator can only achieve long-term impact if it designs and develops a metadata registry and the associated processes and tools to facilitate the integration and correct use and future evolution of the metadata registry in the user community.

Specifically, three elements are necessary to be successful in the creation and evolution of a metadata registry:

- Initial creation of metadata schema. It is desirable that this initial design integrates, to the extent possible, all the metadata schema. It is important that a foundation of shared metadata types for all the metadata schema be created without proliferation of formats for metadata that “means the same thing.” If this not achieved, then the users’ ability to search across collections will be greatly reduced and the situation will be aggravated in the future when the set of metadata schema grows.
- Subsequent creation of new metadata schema must be a controlled process. New schema must, to the extent possible, be developed on the same shared metadata types and be integrated with the existing set of metadata schema.
- A process must be developed for the long-term controlled evolution, maintenance, and use of metadata schema in Curator system. This process may be established after the initial creation of metadata schema, but must be in place before and be used during the subsequent creation of new metadata schema.

Relevant standards that may help us in the Curator project are:

- *ISO/IEC 11179 Specification and Standardization of Data Elements* [23];
- *OASIS Registry/Repository (for XML Schemas)* [24];
- *Preservation Metadata and the OASIS Information Model (OCLC/RLG WG on Preservation Metadata)*. [25]

Obviously, tools should be developed to support the development, evolution and control of the metadata registry.

5.5.3 Metadata Representation Standards

Standards for metadata representation are not mature. Many efforts default to using XML Schema to represent metadata because XML in most cases is the preferred implementation formalism and no additional overhead is involved in creating higher-level metadata descriptions or ontologies. Before committing to XML Schema as a modeling formalist (as distinct from an implementation vehicle), we will revisit the status of efforts on metadata representation standards, including XML Schema [26], RDF [27], Dublin Core [28], XML METS [29], VERS [30], and MOF [31]. We will follow standards representation

development efforts and leverage these when possible.

5.5.4 Extraction of Metadata from Models

Despite the best efforts of developers, software documentation tends to become out of date with respect to the code itself. Often the only recourse to understand a component is to read its code, a labor-intensive practice. It is therefore desirable to apply reverse-engineering techniques to automatically extract a domain model (ontology) from the component. Although in general, this is beyond the state of the art, it may be possible in situations where the domain is already reasonably well defined, as with the Curator. Techniques for performing this analysis are described in [32]. The research question to be explored is the extent to which accurate metadata can be automatically extracted from existing code for Earth system models.

5.5.5 Implications of Abstraction in an HPC Environment

The goal of ESMF and ESG is to provide integrated resources to climate scientists. The primary means for providing such integration is abstraction. That is, an integrated framework must provide generic constructs suitable for being specialized in a variety of ways. While some assert that there is an inevitable tradeoff between abstraction and performance, other contend that abstraction offers many opportunities for optimization that can be invisible to the user, and that this is particularly desirable in the HPC domain. We will examine the results of abstraction in ESMF, ESG and the Curator to see what the effects on performance are.

5.5.6 Automatic Generation of Component Wrappers and Applications

Existing components (which may include gridded components such as land models, as well as couplers and datasets) must be adapted in order to work with ESMF. Some elements of that adaptation, such as writing boilerplate for couplers, are routine and beg for automation. Such a situation suggests the direct application of generative programming technology [33]. That is, given an appropriate specification of the appropriate component, via its metadata, the appropriate adaptations can be made in the form of *wrappers*. A wrapper is a software module that encapsulates a component so that it can be used in situations where it otherwise cannot easily be used – in this case, as part of an ESMF application. One research question to be explored is the extent to which existing generative programming technology can be used to automatically generate wrappers that do not compromise execution efficiency. Another is the extent to which it is possible to generate auto-generate applications that are viable, technically and scientifically, in the climate domain.

6 Participant Qualifications

In this section, we describe the various participants' background in research and development on Curator-related ideas.

Princeton University and GFDL developed the GFDL Flexible Modeling System (FMS) in a 5-year project to place models used for short-term forecasts (hurricane and cloud system modeling), seasonal prediction and interannual climate variability studies, and secular climate change into a single framework. Models based on FMS have been in production now for several years. As our participation in large-scale studies grows, we found it necessary to develop a standardized runtime environment (FRE: the FMS Runtime Environment). The FRE descriptor of a model configuration can be thought of as akin to what we propose to develop as a “model metadata layer” for the Curator. It is architecturally general, but FMS-specific in its details. We hope to use this as a basis for Curator development.

NCAR, MIT and Princeton/GFDL are core participants of the ESMF: the three technical leads of the ESMF project are all participants here. The ESMF provides data structures that we hope will be community standards for describing Earth system model components. These structures will be a key building block for the Curator. The ESMF Component Database captures some fraction of what the Curator is being proposed for, albeit without the formal basis that is an essential stratum of the Curator.

NCAR is a core development, integration, and service deployment site for the ESG effort. ESG provides an initial substrate of model datasets, a schema, and metadata catalogs representing over 50 TB of extant scientific data. NCAR has also played a primary role in developing the metadata and data “publication “ subsystems that serve the ESG environment, along with the associated web and Grid services underneath. ESG has also begun to publish the model codes and initialization datasets, along with metadata, which represents basic steps towards the more comprehensive objectives expressed in this proposal.

Georgia Institute of Technology contributes expertise in the information technology area, specifically component architectures, their specification and realization, metadata definition and maintenance. This includes the automatic generation of wrappers for adapting software components to support a common communication infrastructure. It also includes extensive experience with metadata including database architecture standardization efforts for ANSI/SPARC and the standardization of a communication environment for NASA.

7 Personnel Requests

- GFDL/Princeton University: 0.5 Technical Staff to coordinate activities between modeling groups (MIT, GFDL, CCSM) and play a core role in definition of model metadata layer;
- NCAR: 1 FTE Software Engineer shared between ESMF and ESG, to design and implement database, query, compatibility and auto-generation tools;
- MIT: 1 student, for metadata and tool research and development; and
- GA Tech: faculty summer support plus 1 student, for metadata and tool research and development.

8 Workplan

The following individuals will be responsible for supervising the specified activities.

DeLuca and Middleton Build the Curator prototype relational database based on the convergence of the ESG, the ESMF Component Database, and, as it evolves, the Curator metadata schema. Design and prototype flexible query, compatibility, and auto-generation tools with the guidance and input of Rugaber/Mark research.

Hill Use the MITgcm model and datasets to experiment with evolving Curator tools in order to assess their ease of use, performance, and correct operation. Investigate how the Grid services model relates to the ESMF component model, and interact with ESG to develop a prototype Grid launcher and portal for Curator applications.

Rugaber/Mark Working with GFDL, MIT, and other modeling groups, ESG, and the ESMF Component Database developers, develop a metadata schema for Earth system models and a technical implementation strategy. Investigate and develop a strategy for linking codes with database entries, either through auto-generation or from user specifications. Explore the auto-

generation of “glue” code for wrapping components and assembling applications.

Balaji Research and validate metadata schema for describing Earth system applications based on three key modeling systems, CCSM, FMS, and MITgcm. Direct the comparative study of model metadata. Liaison with metadata efforts in the observational data community. Use GFDL climate models and datasets to experiment with evolving Curator tools and to assess their ease of use, performance, and correct operation.

The sequence of activities will be roughly that outlined in Section 5.4. The establishment of a metadata schema will be the emphasis in the early part of the Curator project, but we will also be interacting early on with potential users to collect use cases and requirements for the database, query, compatibility, and auto-generation tools. Later in the project, when the focus turns more to active prototyping and user experimentation with these tools, the metadata effort will also continue, in order to refine and adjust the schema in response to lessons learned.