A distributed multi-user role-based model integration framework

K. E. Dorow, I. Gorton & D. A. Thurman
Pacific Northwest National Laboratory, U.S. Department of Energy, USA

Abstract

Integrated computational modelling can be very useful in making quick, yet informed decisions related to environmental issues including Brownfield assessments. Unfortunately, the process of creating meaningful information using this methodology is fraught with difficulties, particularly when multiple computational models are required. Common problems include the inability to seamlessly transfer information between models, the difficulty of incorporating new models and integrating heterogeneous data sources, executing large numbers of model runs in a reasonable time frame, and adequately capturing pedigree information that describes the specific computational steps and data required to reproduce results. While current model integration frameworks have successfully addressed some of these problems, none have addressed all of them. Building on existing work at Pacific Northwest National Laboratory (PNNL), we have created an extensible software architecture for the next generation of model integration frameworks that addresses these issues. This paper describes this architecture that is being developed to support integrated water resource modelling in a metropolitan area.

1 Introduction

The use of computational modelling to help answer questions related to environmental issues is a complex process. In most cases, many people are involved, collaborating together on a common set of information using several models chained together to produce the desired results. While frameworks exist to facilitate the integration of models, they tend to focus on a single user environment [4]. As such, a large burden is placed on the individuals involved to share information with their fellow collaborators. Because such sharing takes

Brownfields: Multimedia Modeling & Assessment, G. Whelan (Editor)

place outside of the model integration framework it is leads to difficulties, such as requiring all participants to agree upon a common data format and transform their results into that format. Other examples of difficulties created by such systems include the inability to efficiently control work flow between participants and the inability to accurately capture all of the information necessary to reproduce results.

To overcome these problems, we created a model integration framework that supports multiple types of users and provides tools to help make their collaboration efficient and effective. The remainder of this paper describes this underlying architecture of this new framework and is structured as follows: Section 2 provides a high level description of the architecture. Section 3 examines the Model Connection Framework, which builds upon an existing technology called the Framework for Risk Assessment in Multimedia Environmental Systems (FRAMES) developed at Pacific Northwest National Laboratory (PNNL). Section 4 discusses the Central Repository component and its support for capturing pedigree information. Section 5 discusses an improved method for integrating models into the framework. Section 6 describes the Data Harvester component, which provides tools for seamless data integration from heterogeneous data sources. Section 7 briefly describes the Study Manager user interface and how it caters to different users. Section 8 discusses components for integrating other applications into the framework, specifically applications for analysis and visualization of modelling results. Section 9 presents the Distributed Computing component. Finally, Section 10 discusses the current status of the implementation of the framework.

2 Architecture

In the process of creating the architecture for this framework, we examined the processes and procedures followed by the integrated water resource modelling program at the King County Department of Natural Resources and Parks in Washington State [3]. This program provides hydrological data which is used to help in the decision making process for land / water usage issues in the greater Seattle metropolitan and outlying areas. Based on the needs identified through this examination, the architecture shown in Figure 1 was developed.
The framework architecture is divided into four sections:

- Modelling / Applications—the components in this section support the creation and execution of modelling simulations and provide tools to perform analysis and visualizations of the results.
- Data Management—the components in this section allow for the easy incorporation of data from both internal and external sources and provide a gateway to the information stored in the Central Repository.
- Central Repository—the component in this section provides persistent data storage for the framework.
- Distributed Computing—the components in this section provide a mechanism to allow model executions to be performed across a network infrastructure.

Brownfields: Multimedia Modeling & Assessment, G. Whelan (Editor)

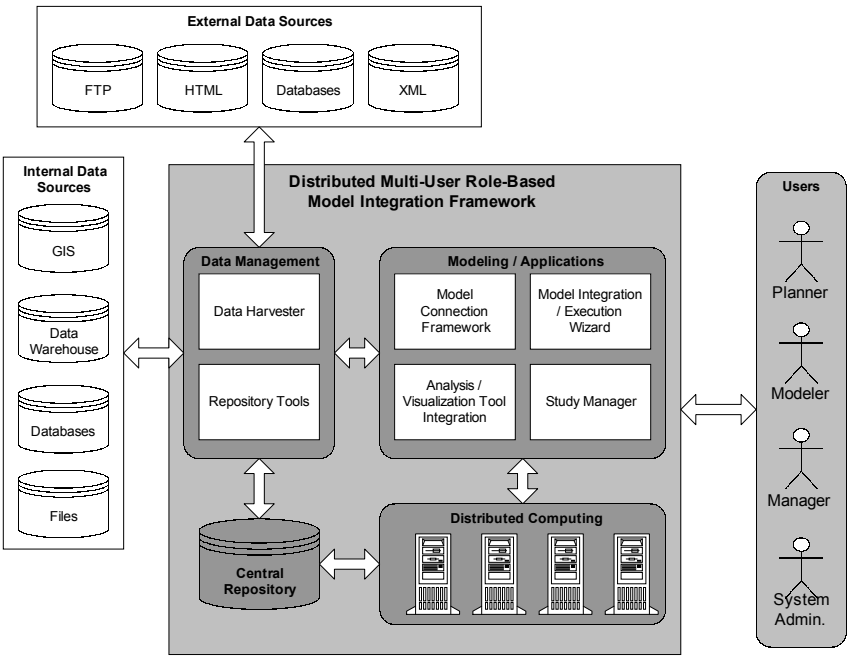

Figure 1: High-level architecture.

In addition, it supports four types of users:

- Manager—a participant who is interested in high level summaries of the results from modelling simulations to aide in the decision making process.
- Planner—a participant who can translate a proposed development scenario into specific changes to the environment (e.g. land usage, water usage, etc.), which can then be simulated as a sequence of model executions (a.k.a. a study).
- Modeler—a participant who performs model calibrations and executions to support studies that have been defined by a planner.
- System Administrator—a participant that maintains the integrity of the software system and performs reporting and archiving tasks.

In the sections that follow, the underlying components within the architectural framework are discussed in more detail.

3 Model Connection Framework

The Model Connection Framework facilitates the seamless transfer of information between the different elements within the modelling simulation. These elements include different types of models (which can be written in

different languages, operate on different scales, and even execute on different computing platforms), data sources (which can vary from well defined databases to structured text files), and analysis applications. Building on extensive research and development already completed in this area, the model connection framework component is being built on top of FRAMES [1]. FRAMES is a platform that allows legacy disparate models, databases, and frameworks to communicate in a plug and play atmosphere. It is a dynamic, reusable software interface structure that:

- Establishes documented data specifications (which include as a criteria a minimum-set data transfer) to allow models and data sources to be connected within the overall software structure.
- Provides a common Application Programming Interface (data protocols) to enable data transfer between models.
- Contains multiple "medium-specific" models (for example, air, water, and human impacts).
- Contains a database of chemical properties with associated environmental parameters.
- Provides a way to perform sensitivity and uncertainty analysis on data from all kinds of deterministic models integrated into FRAMES.
- Retains medium-specific model developer's choice of programming environment (languages, styles, tools) when inserting their software model into the FRAMES structure [1].
- Contains a user interface and modelling tools that are focused specifically for modelers.

In addition to this rich set of features, we are extending FRAMES in the following ways to support the framework architecture:

- Creating a centralized data processing service to support multiple concurrent users working collaboratively on modelling scenarios.
- Adding an automated meta-data capturing mechanism to help keep track of user interactions with the system (who performed what operations on which data and when it was performed).
- Enhancing the data transformation services provided to support new standards in self-defining data formats such as XML (Extensible Markup Language) and associated transformation mechanisms such as XSLT (Extensible Stylesheet Transformations).

4 Central Repository

The Central Repository component is a storage facility for the information that is shared amongst all other components in the framework. It contains all data produced within or imported into the modelling framework. In addition, it acts as a clearinghouse for all models, applications, and data sources that have been

registered for use within modelling simulations. It also captures pedigree data as information is transferred in and out of the repository. Pedigree information is metadata that describes the lineage of the information with which it is associated (where it came from, how it was created, when it was produced). As such, it is critical in providing reproducibility and defensibility of results produced within the framework. In addition, the Central Repository also contains tools to assist a system administrator in maintaining the integrity of the framework, as well as reporting and data archiving utilities.

5 Model Integration/Execution Wizard

The Model Integration/Execution Wizard component simplifies the task of integrating new models into the framework. Existing frameworks, such as FRAMES, require users to write special programs to make models adapt to the framework-supplied data specifications. These special programs, called pre-processors and post-processors, perform the task of converting data between the standard format specified by the framework and the model specific formats. Our approach will allow users to incorporate models without special programs to perform the data format conversions. Instead, graphical tools will be incorporated to allow the user to visually map the model-specific data and control formats to the framework specifications. This mapping process will only be necessary one time, when the model is registered within the framework. The resulting data transformation definitions will then be stored in the repository and are called upon when the given model is employed as part of a simulation.

6 Data Harvester

Computational models typically require data from a wide variety of heterogeneous data sources, including both those that are internal to an organization (databases, flat files, GIS, etc.) and those that are externally available via the Internet (web pages, web services, XML, FTP, etc.). In addition to the wide range of types, the availability of the data sources can be dynamic. Data sources like internal databases may be available all the time and, as such, can support data extraction on demand. Data sources like files on an FTP site, however, may only be available for a limited time and, as such, must be completely imported into the modelling framework. Manually incorporating data from heterogeneous data sources for use in the modelling framework can be very time consuming. Retrieving the data and transforming it to the required framework specifications by hand is cumbersome, error-prone work.

To simplify the process of importing and transforming data we will create the Data Harvester component to provide tools to manage connections and data storage for all different types of data sources. In addition, it will support a wide variety of data source availabilities (direct connections, scheduled extractions, etc.). It will also make the process of transforming the data to the required framework specifications simpler by providing graphical tools that allow a user to create the transformations visually.

Brownfields: Multimedia Modeling & Assessment, G. Whelan (Editor)

7 Study Manager

The Study Manager component provides a collaborative user interface that caters specifically to the non-modeler participants of the framework (managers and planners). It contains a map-based interface that allows the user to generate an integrated modelling simulation by selecting regions of interest from a map (or series of maps). It also has a workflow management system that can be used to break up an integrated modelling simulation into specific tasks (e.g. data collection, model calibration, model execution, results analysis) that can then be assigned to the appropriate participants in the system. The workflow management system can also track the progress of these tasks and display status information as they are completed. See [2] for a more detailed description of the development of this component.

8 Analysis/Visualization Tool Integration

An important aspect of integrated modelling is the ability to take the raw data output from models and interpret its meaning either by further analysis and / or through visualization. The Analysis/Visualization Tool Integration component facilitates this aspect by allowing users to easily export modelling results to 3rd party tools (e.g., MATLab, GIS applications, TechPlot, Excel). Like the Model Integration Wizard, it provides the ability to graphically describe the transformation to map the model output data specifications to a format required by a particular 3rd party tool. It also provides tools to define the specifics of a 3rd party tool such that it can be launched directly from within the modelling environment. A reporting capability is also included so that users can track the content and usage of the integrated modelling framework.

9 Distributed Computing

The Distributed Computing component provides a mechanism to allow modelling execution tasks to be processed on available computing resources across a network infrastructure. It gives the modelling framework the ability to take advantage of economies of scale (as more computers are added to the network infrastructure, they automatically become available to all users). Modelling simulations that contain multiple, independent model execution tasks can greatly benefit from this functionality by distributing those tasks across different computing resources, drastically reducing the time necessary to get results. It also provides the ability for models that execute on varying platforms (Windows, Linux, Unix, etc.) to be incorporated into the model integration framework seamlessly.

10 Status

The framework as described is currently being implemented for deployment at the King County Department of Natural Resources and Parks. By the time of

Brownfields: Multimedia Modeling & Assessment, G. Whelan (Editor)

publication of this paper, a demonstration prototype containing the Model Connection Framework, Central Repository, and Study Manager will be completed. The entire framework is scheduled to be completed and deployed to King County in December of 2005.

References

[1] Whelan, G., Castleton, K.J., Buck, J.W., Gelston, G.M., Hoopes, B.L., Pelton, M.A., Strenge, D.L., and Kickert, R.N. Concepts of a framework for risk analysis in multimedia environmental systems (FRAMES). PNNL-11748. Pacific Northwest National Laboratory, Richland, Washington. 1997.

[2] Thurman, D.A., Cowell, A.J., Taira, R. and Frodge J. Designing a collaborative problem solving environment for integrated water resource modeling. These proceedings. Siena, Italy 2004.

[3] Thurman, D.A., Peterson, T.S. and Frodge, J. Defining requirements for an integrated water resource modelling system. In *Proceedings of the 6th World Multi-Conference on Systemics, Cybernetics, and Informatics, vol. XVIII.* pp.474-479. Orlando, FL. 2002.

[4] Taira, R. Y., Johnson, D.M. and Thurman, D.A. An evaluation of model integration frameworks for the integrated water resource modelling system. PNNL-. Pacific Northwest National Laboratory, Richland, WA. 2003.

Brownfields: Multimedia Modeling & Assessment, G. Whelan (Editor)