THE EARTH SYSTEM GRID DISCOVERY AND SEMANTIC WEB TECHNOLOGIES

Line Pouchard
(Oak Ridge National Laboratory)
pouchardlc@ornl.gov
Luca Cinquini, **Gary Strand**
(National Center for Atmospheric Research)

The Earth System Grid [1] project includes many active participants. The senior staff are:

Ian Foster, Veronika Nefedova (Argonne National Laboratory), Line Pouchard, David Bernholdt, Kasidit Chanchio, Meili Chen (Oak Ridge National Laboratory,) Dean Williams, Bob Drach (Lawrence Livermore National Laboratory), Luca Cinquini, Don Middleton, David Brown, Peter Fox, Jose Garcia,[3] Gary Strand (National Center for Atmospheric Research,) Arie Shoshani, Alex Sim (Lawrence Berkeley National Laboratory,) Shishir S. Bharathi, Ann Chervenak, Carl Kesselman (University of Southern California Information Science Institute.)

ABSTRACT

The Earth System Grid (ESG) [1] is developing a virtual environment based on Grid technologies for the earth sciences and others analysizing the impacts of global climate changes. The goal of ESG is to provide secure access to data and data transfer to scientists and interested communities. Data discovery through the use of metadata has become a major focus of ESG. Metadata schemas, a prototype ontology [2], search and discovery services have been developed and are continuously being improved. ESG discovery mechanisms are being deployed through a service architecture. ESG requirements for data discovery are discussed for the current and projected amount of data in the earth sciences. The need for semantics in ESG and ESG services are also discussed in the context of the Semantic Web. The paper concludes with remarks on what Semantic Web technologies and scientific Grid projects such as ESG may leverage from each other.

KEYWORDS

Grid Services. Semantic Web. Ontologies. Data Discovery. Earth sciences.

1 INTRODUCTION

In emerging grids and Grid Computing, shared, distributed, and heterogeneous computing and data resources enable scientific advancement through collaborative research and collaboratories. One goal is to provide scientists with seamless, reliable, secure and inexpensive access to resources typically out of reach for many [3][4]. The management of these resources is complex, time-consuming, and not subjected to a centralized control. In data-intensive scientific domains, such as the earth sciences, high-energy physics, and astronomy, terabytes of data will be acquired from simulations performed on supercomputers across the nation and abroad. Helping scientists to efficiently search and retrieve information, manage data, record their observations, and generally perform logistics tasks associated with the pursuit of science is crucial due to the increasing volume of data produced in these domains.

The Earth System Grid (ESG) is developing a virtual collaborative environment based on Grid technologies such as Globus tools [5] to facilitate analyzing the impacts of global climate change at national laboratories, universities and other laboratories [6]. ESG is a project of the U.S. Department of Energy Scientific Discovery through Advanced Computing program. ESG will provide access to data produced by earth and climate science simulations through a Web portal. Through that portal climate scientists and researchers utilize distributed resources to discover, access, select, and analyze model data produced and stored in archives. The challenges posed by the volumes of data stored, the issues surrounding secure access and the choice of resources require smarter and increasingly flexible tools. Some of the technologies developed for the Semantic Web may prove very useful for proposing some solutions to the challenges outline above.

This paper presents some requirements for searching and retrieval of scientific information and how ESG uses metadata services for answering these requirements. The importance of data quality and provenance of datasets increases in distributed datagrids. With data sizes and distribution, it is difficult to determine where a file comes from, what transformations it went through and who produced it, and what variables were used for the obtention of a dataset. This ancillary information may include things that are as unrelated as simulation codes or the sponsor's name.

The paper also discusses a prototype ESG ontology that distinguishes between metadata specific to the Earth Sciences and metadata common to grid projects.

2 ESG REQUIREMENTS

User requirements were established by close collaboration between computer scientists and domain experts. Of concern to the Semantic Web is the need of scientific users to search and retrieve files and collections of files located in mass archives, how provenance is established, and the potential role of ontologies. Searches are expected to point to datasets based on search criteria such as date and time coverage, presence of specified variables, type of simulation models used for a particular dataset, and related datasets. Access through a single point of entry from a scientist's desktop is required.

Users are climate scientists at national laboratories, other government agencies, and universities around the country and abroad where earth scientists do research. A motivation for the development of ESG is to bring online resources to users with limited access to community data. Users need to move very large datasets between sites that have sufficient computing power and simulation software to run the models for analysis. Data transfer may be initiated from a third site, and from a desktop machine. Because of the size of datasets, scientists want to know the "content" of a dataset before deciding to transfer. Others want to store their data in the archives and make it available to the community. Another advantage is avoiding duplication such as reprocessing simulations several times by different users because they do not know that an existing model and results already exist. The importance of avoiding reprocessing comes from the fact that these simulations run for one to several weeks, consume many compute and team/hour resources.

Data that ESG handles is simulation data produced by running climate simulation models. ESG is not currently expected to manipulate raw data outputted by observation stations. ESG data is processed data, i.e. data that has been obtained through simulations and modeling. Data already processed is used to create new models with the effect that at the end of a chain, at the previous , it may be difficult to determine what processdata a dataset of a particular model may be obtained from other model datasets. Some datasets are linked to each other by model configuration, and some datasets are part of collections. Data sizes already are barely manageable and data loss will occur if discovery mechanisms are not soon and greatly improved. As of July 2003, the estimated total volume of data to be created by running the necessary simulations for the Inter-governmental Panel on Climate Change is 18.91 Terabytes.

3 SEARCH AND RETRIEVAL USE CASES

Search and retrieval of datasets are a primary functionality of ESG (the others are data transfer and security of all transactions). The ability to locate and obtain datasets as easily and seamlessly as possible is crucial to climate scientists, and therefore to ESG. The time currently needed for locating a file must be shortened and the human input automated.

In ESG search and retrieval are based on an ESG metadata schema that will evolve as new metadata becomes necessary. Fine granularity of users and operations they need to perform on ESG were defined and was essential for designing the ESG metadata schema. For search and retrieval, these scenarios are described here in broad terms. (1) An ESG publisher is a human or machine creating datasets, annotating datasets with metadata and submitting them to ESG for publication. (2) A publisher extracts it from the ESG metadata schema and assigns it to his data. (3) An authenticated ESG user browses and/or searches ESG metadata catalogs for selecting datasets to download. This ESG user accesses datasets of interest and transfers it from an archive to another site. (4)An ESG user as computer application creates the necessary metadata as datasets are produced. The user application sets up the new metadata file identifying unique coordinates and time describing the data.

Some examples of complex queries include (1) identify datasets containing such and such variables across datasets with unrelated schemas. --currently, a query for datasets containing variables must include a time range.-- (2) return slices of data for files containing the variables "wind" and temperature" at these geospatial coordinates. Slices of data would return only the "piece" of a dataset containing the above variables, not the whole dataset containing

them with irrelevant information to the particular experiment; (3) return the datasets above from data archives held in world repositories such as the UK, Japan, continental Europe. Ideal cataloging and discovery scenarios for climate scientists include the automatic generation of metadata catalogs, transparent access regardless of the archive location, searches allowing discovery through multiple catalogs based on different metadata schemas and the extensibility of these catalogs.

ESG is required to capture model run descriptions (including input scenarios, time period), model configuration information (such components like atmosphere, ice, ocean), identification of input datasets, pointers to documentation, sites where the models are run, and people who carried out the model integration and submission to archives. A very important requirement for ESG was to capture model experiments related by "parent," "child," and "sibling" datasets. For instance, it is important to identify if a given dataset was produced by a model run that is child of another model run.

4 THE NEED FOR SEMANTICS

Grid architectures have been somewhat service-oriented and emphasize operations that can be performed on data using the associated metadata schemas if they exist, rather than developing metadata that help identifying datasets and their location. Scientists typically know and trust what to expect from a particular dataset at a particular archive, and login to this archive to perform analytical operations. Provenance is also known this way. This information is often held in a scientist's brain and/or in the data manager's administering the archive. It may also be held in static catalogs not located or linked to datasets, within the dataset itself, therefore inaccessible before downloading a file. Catalogs, when they exist, are specific to formats and conventions in which output results are produced.

Metadata for scientific information is any information scientists may need or want when they make decisions about actions to perform on data available for their research. This "ancillary" information has always been important and available from multiple sources, including stand-alone collections, lab notebooks, heterogeneous online sources. Information about the design of an experiment, experimental conditions and parameters may be contained in a lab paper or electronic notebook. Information about the data such as its time periods, versions, what variables are included may be stored with the data itself, so that the only way to access it is by examining file content

Metadata in grid architectures is often implicit, and contained within each service but not described. Metadata schemas are found in database tables and storage systems that are not usually directly accessible to a scientific user and may be limited for discovery purposes. This state of things makes metadata difficult to access and compare. Such metadata contains little semantics beyond an entity-relationship model, the services designed to use it are not interoperable unless by preliminary design and not composeable. (Services can be composed when they are not simply invoked by methods through APIs or remote calls but become parts of more complex, high-level "conversations," where content may be based on ontologies [7].) At best, metadata is described in XML with a data dictionary, an entity relationship model, or a database schema are available. Redundancy, overlap, and gaps may occur without the explicit knowledge of the user, leading to interpretation errors. By expressing relationships between metadata elements and increasing inter-operability between earth science metadata, ontologies attempt to remove some ambiguity.

In several disciplines, large scientific projects are the product of teams of collaborators at multiple institutions. Some information like scientific use may be specific to a project. But the need for access to people, projects, data provenance, storage areas, and security are common across the board. In addition to granularity in ontologies, criteria for separating common concepts and discipline or project specific concepts are needed.

5 ESG PROTOTYPE ONTOLOGY

A prototype ontology for the earth sciences has also been developed using Protégé-2000 [8]. This ontology specifies broader categories for kinds of information found in ESG and other Grid projects and is described in details in [2]. The ESG ontology is based on the ESG metadata schema. Pursuing collaboration with the British National Environmental Researh

Council (NERC) in the UK has provided motivation for this ontology. After determining some similar requirements, ESG and NERC have planned to leverage some tools and schemas from each other.

Metadata important for both teams appear similar in content but the paradigms under which they are organized are different. As information is expressed differently and classified using different, overlapping or gaping categories, it is harder to relate schemas and share tools. The ESG schemas describe entities while the NERC schemas describe processes. The ESG classification system centers on capturing (static) information concerning data (formats, variables, collections) and people. NERC focuses on usage and discovery of datasets, and classifies its metadata accordingly. This disparity increases the complexity of the relationships between schemas because one-to-one mappings between two schemas must be augmented by conditional rules to be accurate. If conditional rules are not set, mappings may introduce errors not detectable to the scientist. Schema entities overlap in their "meaning" so that one entity in one schema may refer to data annotated by several entities in the other schema. Conversely, one entity in a schema may refer only to some instances described in the other schema's corresponding entity. If a schema is changed, the mapping must also be changed. This becomes non-trivial when a project uses N2 mappings to many schemas rather than N mappings to an ontology.

The ESG ontology contains the disjoint classes of Pedigree, Scientific Use, Datasets, Services, Access, and Other. The ESG Pedigree represents the line of ancestry for a collection of individual files or a single file. People associated with a dataset such as PI, Publisher, and other roles are part of Pedigree. Provenance is also a slot in pedigree and records names or IDs of datasets that served as input or output for a particular dataset simulation. Some pedigree information uses the Dublin Core. The Scientific Use class specifies all information that is pertinent to the use of a dataset and its production. ESG Scientific Use is likely to be of great interest to a scientist. It describes model configurations, initial boundary conditions, model version, time and space coverage, measurement ranges and units. The Scientific Use subclass "Investigation" categorizes ESG experiments in "campaign", "ensemble", "observation" and "analysis." Datasets include "collections", "aggregations", and "parameters." "Access" refers to security information and "Other" to (largely) manual comments and references.

Ideally, with Provenance and Scientific Use, a user would trace the conditions under which a particular dataset has been constructed, its input, along with the line of ancestry from the dataset is derived. Ultimately some simulations would not need to be run several times for obtaining analysis results. Provenance and Scientific Use may help build trust in data and allow use of a larger amount of datasets. Currently trust largely depends on a scientist knowing another, publications, and an institutional source. Provenance may also be used for verification by the scientist who himself produced data at an earlier stage rather than performing frustrating searches in old notes.

In ESG, a service associates climate data formats with servers capable of processing them (Figure 1).

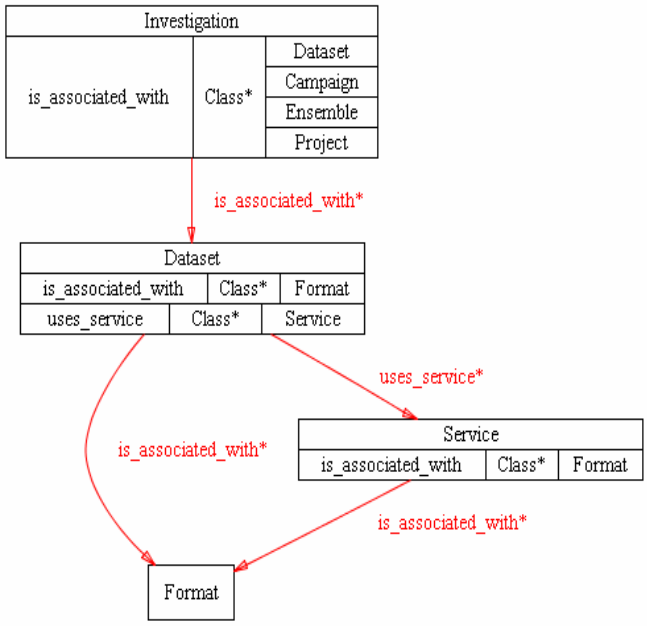

Figure 1: ESG Service Class.

A type of processing may be treated by several servers, and a server may be able to treat several types of formats. The ESG metadata schema also includes the ability to assign a standard format or convention to the data format attribute. There are several dataset formats corresponding to the simulation models run. These formats are typically unrelated in syntax and semantics.

Relationships in the ESG ontology include (Figure 2):

Is_part_of: a dataset is_part_of an investigation.

Is_parent_of and the inverse_of_parent, is_sibling_of: simulationA is_parent_of simulation B, simulation B is_inverse_of simulation B.

Is_set_of: Ensemble is_a_set_of simulation.

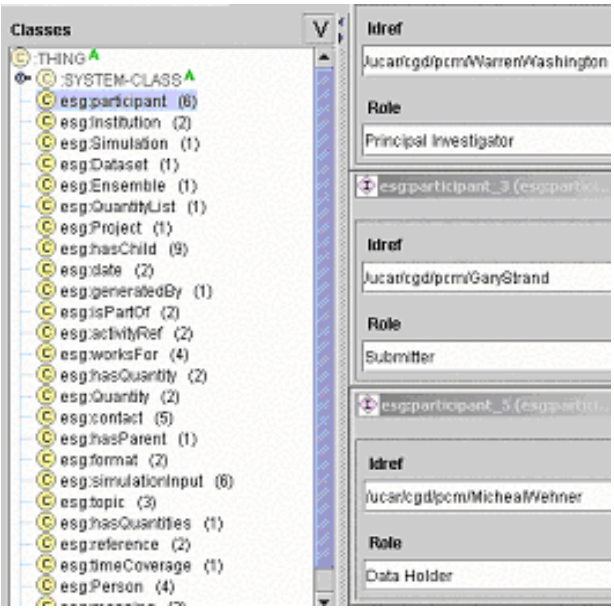

Figure 2: Some ESG classes and relationships

The logical separation of classes between what is used in ESG and what may be used in other grid projects has been a leading principle in building the ESG prototype ontology. While Scientific Use and Investigation are domain-specific, Access, Dataset, and Pedigree may be common to several grid projects under certain conditions based on rules. However, these high-level ontology classes lose some applicability depending on the required granularity. For instance, while the sub-classes of Scientific Use such as Investigation, Experiment, Observation may apply to other Grid project, Campaign and Ensemble may not. Dataset slots such as associated project, and dataset owner may be common, but not archive locations and parameters. Required granularity will depend on the expected use of ontologies in Grid projects. As tools suitable to several projects such as Globus libraries, metadata catalog services and replica location services become more common, some metadata in these catalogs may be re-used based on ontologies across different grid projects. Ontologies for "logistics" or "house cleaning tasks" information is an example.

6 ESG SOLUTIONS (IN PROGRESS)

ESG developed its own XML metadata schema specifically formatted for earth sciences modeled data but extensible. ESG evaluated several existing data description mechanisms for use with earth sciences data. The Dublin Core was not rich enough to support scientific data although it may be used for people and project metadata at a later stage. The earth sciences ISO standards proved too complex for ESG purposes. Two key ESG contributions towards discovery services are the representation of collections, and the implementation of Logical and Physical file names (for a better term).

Collections

The ESG schema focuses on describing collections of files and search and discovery of collections of files. Files may be assembled in an ESG collection by a user (for instance, myCollection). A collection may consist of a single file, several files, and other collections. These collections and the inclusion criteria may be available to other users or not. Criteria for building collections include relations between files such as parent, child, sibling relationships. Any other relations between files that are of interest to a user or collection builder may also constitute the criteria for inclusion. For instance, collections based on a geographical, dimensional, time-coverage are possible. Collections assembled based on model parameters are also possible.

File Names

ESG uses logical file names to reference a collection and physical file names to locate it. A query to the ESG discovery services returns logical file names according to search criteria. The logical file may represent a single file or a set of logically related files such as a collection. The logical file name of interest points to a set of physical files, possibly in different archives, for the user to chose from. The user then chooses a location from where to download the file or collection.

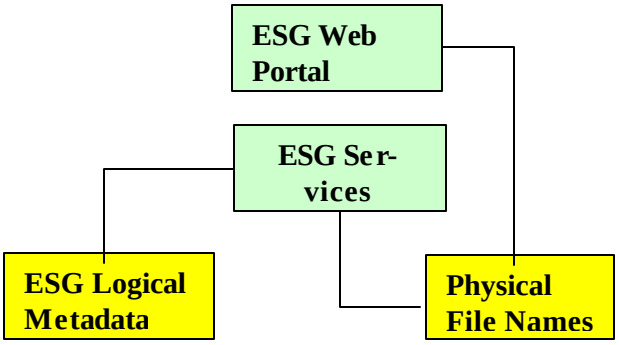

Figure 3: ESG Discovery

ESG services may only be partially understood in the context of the Semantic Web. ESG Services are static and persistent and cannot be composed to accomplish a goal. Workflow is pre-defined and services do not exchange service-based information based on semantics. The ESG prototype ontology attempts to clarify relationships between datasets, scientific use of data, and data pedigree for the ESG metadata schema. Metadata catalogs for data discovery are currently contained in object databases and must be manually updated with each new metadata class. ESG metadata services do not play the role of a service broker or coordinator for the ESG Logical Metadata Catalog service and the Physical filename service.

7 CONCLUSION

An expected development for ESG is the availability of a user-friendly annotation tool to minimize data entry. Another projected development is a tool for the automatic or semi-automatic extraction of metadata. A legacy problem exists for inserting metadata in files already present in the archives. This is currently done by hands for a few test files. Improvements will be made to better track and record provenance and data transformations as this is currently incomplete and many gaps and overlaps. Better tracking provenance may require expressing parts of the ESG schema and relationships in a description language such as RDF.

The Earth System Grid provides the Semantic Web and Semantic Grid real life complexity and applications for testing the limits on some of its capabilities. For example, Semantic Web services such as ontology based annotations may consider requirements for tools performing automatic or semi-automatic annotations. Web Service composition will be tested for grid complexity and scalability.

The Semantic Web efforts have highlighted the need for interoperability based on content, and started offering tools toward this goal. It may bring to projects like ESG a more flexible approach for designing schemas with relationships, extensibility, and interoperability. In particular a more expressive and stable representation language such as RDF starts being accepted in the Grid communities. Methods for partial mappings and ontology reconciliation using pieces of common, small ontologies already exist and could be adapted for Grid purposes.

Interdisciplinary collaborations and the number of participants in scientific projects may only increase. The Semantic Web's focus on mechanisms for sharing information based on content, and tools for handling the complexity may bring a measure of relief to current obstacles in scientific grids.

REFERENCES

[1]The Earth System Grid. http://www.earthsystemgrid.org/.

[2] An Ontology for Scientific Information in a Grid Environment: the Earth System Grid, Line Pouchard, Luca Cinquini, Bob Drach, and the ESG team, Proceedings of the Symposium on Cluster Computing and the Grid (CCGrid 2003), Tokyo, Japan, May 12-15, 2003

[3]The Anatomy of the Grid: Enabling Scalable Virtual Organizations. Ian Foster, Carl Kesselman, Steven Tuecke. International Journal of High Performance Computing Applications, 2001. 15 (3): p. 200-222

[4] The Grid: Blueprint for a new Computing Infrastructure. Ian Foster and Carl Kesselman, eds. Morgan Kaufmann, 1999.

[5]The Globus project. http://www.globus.org/.

[6] The Earth System Grid: Turning Climate Datasets into Community Resources. AMS 2002.

[7] Web Services and Beyond. Munindar P. Singh and Michael N. Huhns.] Forthcoming.

[8] Protégé-2000, http://protégé.stanford.edu/.

The submitted manuscript has been authored by a contractor of the U.S. Government under Contract No. DE-AC-05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow other to do so, for U.S. Government purposes.