The Earth System Grid Discovery and Semantic Web Technologies

Line Pouchard (Oak Ridge National Laboratory, Oak Ridge, TN 37831-6367, pouchardlc@ornl.gov)

Luca Cinquini, **Gary Strand** (National Center for Atmospheric Research, Boulder, CO, 80301, luca@ucar.edu, strandwg@ucar.edu)

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

Abstract

The Earth System Grid (ESG) is developing a virtual environment based on Grid technologies for the earth sciences and others analyzing the impacts of global climate changes. The goal of ESG is to provide discovery and secure access to very large datasets for earth sciences research. Data discovery through the use of metadata has become a major focus of ESG. Metadata schemas, a prototype ontology, search and discovery services have been developed. ESG discovery mechanisms are being deployed through a service architecture. ESG requirements for data discovery, the need for semantics, and ESG services are discussed. The paper also discusses relevance to the Semantic Web in the context of Grid computing serving pre-defined scientific communities.

Keywords

Grid Services. Semantic Grid. 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 (Foster, et. al. 2000, 2001). 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 (The Globus Project) to facilitate analyzing the impacts of global climate change at national laboratories, universities and other laboratories (Middleton). 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 answers these requirements by implementing metadata services. The importance and challenges posed by data quality and provenance of datasets increases in distributed data grids with sizes and multiplicity of users. Tracing provenance, a concept that loosely describes where a file comes from and what transformations it went through, becomes crucial. This ancillary information may include things that are as unrelated as names of simulation models 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. Scientists providing expertise for the Inter-governmental Panel on Climate Change (IPCC – the Kyoto treaty) are a main target user group. 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 computational 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, it may be difficult to determine what process a dataset went through. 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 Intergovernmental Panel on Climate Change is 18.91 Terabytes.

3 Search and Retrieval Use Cases

Search and retrieval of datasets generated by earth science simulations 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. At the time of this writing, the last scenario has not been implemented.

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 like ESG are service-oriented and emphasize operations performed on data such as high-speed transfer, mass storage and security, rather than content descriptions and annotations that help characterize the data. Scientists typically know what to expect from a simulation model and trust known simulation data producers and storage sites as they have always done. However, this method is no longer practical and reliable due to the

size and multiplication of simulation datasets produced by the newest supercomputers. In a non-grid environment users login to a site by client-server protocol, and transfer data to a convenient location to perform analytical operations. Provenance of a dataset, the information that indicates where a dataset comes from and what models were used in its production is known in an ad hoc fashion: the information is held in a scientist's brain and/or in the data manager's who administers the archive. Lists of datasets may be contained in electronic catalogs with little known information beyond the filename. (In ESG, some file names indicate the name of the model, the area of study, for instance atmosphere, and the dataset format). Information such as variables and dimensional coverage is contained within the data itself, therefore inaccessible before downloading a file.

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 personal files, lab notebooks, heterogeneous online sources, and human memory. Information about the design of an experiment, experimental conditions and results may be contained in a published paper and shared or not with other researchers. 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 for grid data is often implicit, and sometimes used within a grid service but not described. Some 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 and available in XML with a data dictionary. 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 interoperability between earth science metadata, ontologies attempt to remove some ambiguity.

In many disciplines, scientific projects are the product of teams of collaborators at multiple institutions. The need for access to people, projects, data provenance, storage areas, and security are common across the board. In addition, criteria for separating concepts common and discipline or project specific concepts are needed. The ESG prototype ontology was developed with this goal.

5 ESG Prototype Ontology

The ESG ontology was developed using Protégé-2000. It specifies broader categories for kinds of information found in ESG and other Grid projects and is described in details in (Pouchard, et. al. 2003). The ESG ontology is based on the ESG metadata schema. Pursuing collaboration with the British Atmospheric Data Centre (BADC) of the National Environmental Research Council (NERC) and the CCLRC e-Science Centre in the UK has also provided motivation for this ontology. After determining some similar requirements, ESG and NERC have planned to leverage some tools and schemas from each other, but no mapping of schemas has been envisioned at this point.

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 lack of congruence 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 sub-class "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.

With Provenance and Scientific Use information, a user may trace the conditions under which a particular dataset has been constructed, the simulation input, and the line of ancestry from where the dataset is derived. This reliable information will save the need to re-run simulations several times for obtaining analysis results. Provenance and Scientific Use may help build trust in data and allow re-use of a larger number 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 earth science data formats with servers capable of processing them such as the Distributed Oceanographic Data System, the Live Access Server, etc…) 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 1):

- isPartOf: a dataset is part of an investigation.
- hasChild is the inverse relation of hasParent.: Dataset z has Parent Dataset A (where Dataset A is a collection).
- isDerivedFrom: Dataset A is derived from Dataset C
- generatedBy.: Dataset L is generated by Simulation Model M.

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.

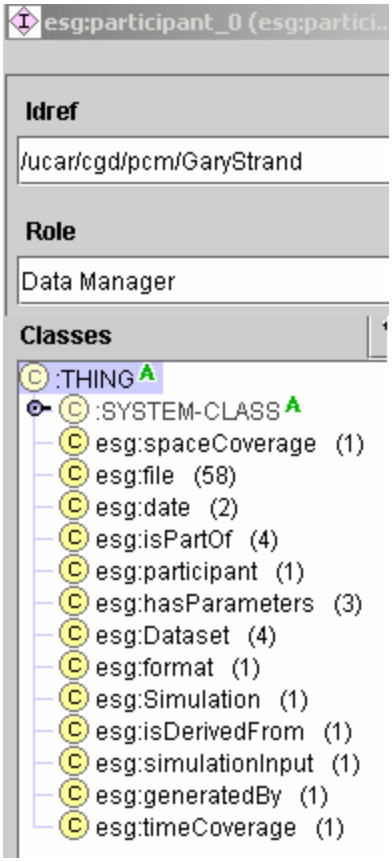

Figure 1: ESG Ontology Classes.

For instance, while the sub-classes of Scientific Use such as Investigation, Experiment, and 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 are an example.

6 Some ESG Solutions and Discussion

ESG developed its own XML metadata schema specifically formatted for earth sciences modeled data. ESG evaluated several existing data description solutions for use with

earth sciences data. The Dublin Core was not rich enough to support scientific data because its primary purpose is to describe paper and electronic publications (such as web pages). It may be used for person-related information. Earth sciences data ISO standards (ISO proved too complex for ESG purposes and timely implementation. Two key ESG contributions towards discovery services are the representation of dataset collections, and the implementation of Logical and Physical file names (for lack of 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) at the time of data publishing or at the stage of discovery. The ESG schema was developed in part to address the requirement of representing collections of files. 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, and 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 multi-dimensional coordinates or time-related coverage are possible. Collections assembled based on model runs (ensembles) are a major component of collections and warrant a separate ontology category.

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.

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.

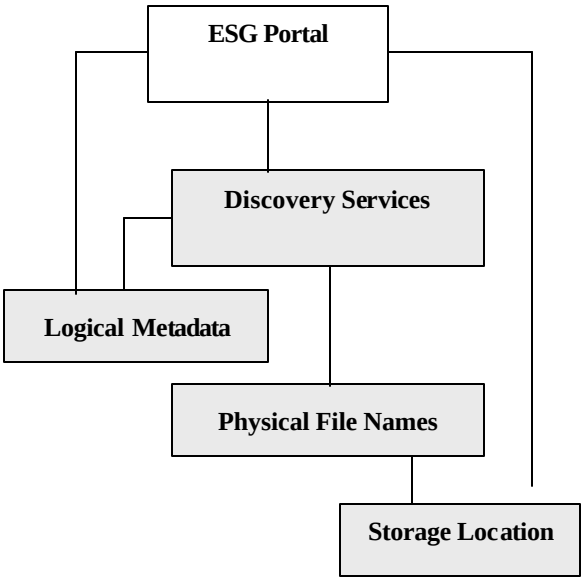

Figure 2: ESG Discovery Architecture.

Discussion

In ESG, users are not expected to access and compare datasets from other, unknown sources, as is the case for Web users. However, within ESG, types of data produced by several families of models must be available for searching, browsing and retrieving from a unique web portal based on metadata information.

The services powering the portal must allow a scientific user to perform operations across a very large amount of distributed data with a few clicks. It is not acceptable for instance that the user repeats searches across collections, storage sites, and model families to find suitable datasets. Given the restricted user community, it is practical to develop a metadata schema for ESG, and provide mappings to existing schemas if necessary. The ESG ontology has provided a framework for developing the ESG schema. The iterative work of detailed concept definitions and the rigor needed for determining relations between entities required in ontology authoring has served well to improve the schema.

For information discovery the Semantic Web may be understood as serving loosely defined communities formed by the nature of and at the moment of their search. For

instance a Web user may search for an ontology needed to construct a Web page (this would define her as belonging to one community). Then this user searches travel information and reservation and accesses a site powered by composeable Web services based on semantics. This would place the user in another community (travelers). ESG efforts pertain to the Semantic Web because it helps define and serve a relatively small and exiting community of users (earth scientists). However, the community served by ESG is much more persistent than the communities described above. The needs of this community are easier to define since they tend to persist, but challenges posed by this community are more daunting,

Another important aspect of ESG and its relationships with Semantic Web technologies to keep in mind is that ESG was designed for a small, precisely targeted community of users. One primary concern was to enable these users to rapidly access, search and retrieve binary datasets from very large archives. These datasets have been poorly annotated. In spite of the state-of-the-art resources involved for producing (supercomputers and simulations) and storing them, they will soon become what amounts to being lost for all practical purposes because of their sheer size and volume. Because of this, ESG is more a part of the Semantic Grid, rather than the Semantic Web.

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.

Acknowledgements

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.

References

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

[2] Pouchard, L., Cinquini, L., Drach., et. al. An Ontology for Scientific Information in a Grid Environment: the Earth System Grid. In *Proceedings of the Symposium on Cluster Computing and the Grid (CCGrid 2003)*. Tokyo, Japan, May 12-15, 2003.

[3] Foster, I., Kesselman, C., Tuecke, S. 2001. The Anatomy of the Grid: Enabling Scalable Virtual Organizations 15 (3): 200-222.

[4] Foster, I., Kesselman, C., eds. 2000. *The Grid: Blueprint for a new Computing Infrastructure*. San Francisco, Calif.: Morgan Kaufmann, Inc.

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

[6] Middleton, D. et.al. The Earth System Grid: Turning Climate Datasets into Community Resources. AMS 2002.

[7] Singh, M. Huhns, M. *Web Services and Beyond.* Forthcoming.

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

[9] British Atmospheric Data Centre. http://www.badc.nerc.ac.uk .

[10] CCLRC e-Science Centre. http://www.esc.dl.ac.uk/.