

An Ontology for Scientific Information in a Grid Environment: the Earth System Grid.

Line Pouchard,¹ Luca Cinquini,³ Bob Drach,⁴ Don Middleton,³ David Bernholdt,¹ Kasidit Chanchio,¹ Ian Foster,² Veronika Nefedova,² David Brown,³ Peter Fox,³ Jose Garcia,³ Gary Strand,³ Dean Williams,⁴ Ann Chervenak,⁵ Carl Kesselman,⁵ Arie Shoshani,⁶ Alex Sim⁶

[1] Oak Ridge National Laboratory, [2] Argonne National Laboratory, [3] National Center for Atmospheric Research, [4] Lawrence Livermore National Laboratory, [5] University of Southern California Information Science Institute, [6] Lawrence Berkeley National Laboratory.

Abstract

In the emerging world of Grid computing, shared computational, data, other distributed resources are becoming available to enable scientific advancement through collaborative research and laboratories. This paper describes the increasing role of ontologies in the context of Grid computing for obtaining, comparing and analyzing data. We present ontology entities and a declarative model that provide the outline for an ontology of scientific information. Relationships between concepts are also given. The implementation of some concepts described in this ontology is discussed within the context of the Earth System Grid II (ESG)[1].

Keywords

Ontology, ontologies, Grid computing, earth sciences, climate.

1. Introduction

In emerging grids and Grid Computing, shared computing and data resources, and other distributed resources are available to enable scientific advancement through collaborative research and laboratories. One goal is to provide scientists with seamless, reliable, secure and inexpensive access to resources typically out of reach for many [2][3]. The management of these resources is complex, time-consuming, and not subjected to a centralized control. For data-intensive scientific domains, such as the earth sciences, high-energy physics, and astronomy, terabytes and soon petabytes of raw data is being acquired from observation and simulation. The potential knowledge contained in this data will be extracted efficiently if scientists can concentrate on “doing real science” rather

than operating complex computer systems to produce results. 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. These tools also need to be customized and integrated in domain-specific contexts.

This paper addresses the increasing role of ontologies in the context of Grid Computing for scientific applications. In a grid environment, information structured in ontologies may become crucial to many operations necessary to obtain and analyze the desired data. For instance, a user may build on the fly a collection of data files based on attributes defined in the ontology. Then files associated in that collection move from their storage place to a desired location but the user may not know the physical location, the name of each individual file, or the names of attributes for the files in the collection. Another example is the selection of a “slice of data” contained in a file or collection based on attributes in the ontology. At a higher level of interoperability, shared ontologies between different systems, and mappings of a domain ontology onto a service are important components of a service-based open architecture and re-use of tools on a semantic basis.

Examples are given for the Earth Sciences based on the efforts of the Earth System Grid, a project of the U.S. Department of Energy Scientific Discovery through Advanced Computing (SciDAC) program. The Earth System Grid II (ESG) is developing a framework that integrates Grid technologies (including the Distributed Oceanographic Data System servers [4], Globus Tools [5], Data Grid technologies and others) to facilitate analyzing the impacts of global climate change at national laboratories, universities and other laboratories

[6]. ESG supports a collaboratory and a Web portal in which climate scientists and researchers may utilize distributed computing resources to discover, access, select, and analyze model data produced and stored on a daily basis on supercomputers across the US.

Motivation for expressing concepts and relationships found between ESG metadata elements in an ontology has come from pursuing a collaboration with the National Environmental Research Council (NERC) Data Grid [7] and the CLRC e-Science Center, UK [8]. ESG and NERC are developing metadata schemas, tools, and access mechanisms for the Earth Sciences communities, while the CLRC is developing such schemas, tools and access mechanisms for a wider variety of scientific disciplines in the UK. Within the earth sciences, disciplines with an interest in these tools include: atmospheric sciences, climate modeling, oceanography, geographical information systems, and meteorology. Sharing data sets and searching across data sets held in both communities and described by different schemas, and re-use of tools is envisioned. Ontologies become crucial to access, browse and perform searches across metadata schemas.

2. Background

An ontology is an explicit specification of a conceptualization where definitions associate concepts, taxonomies, and relationships with human-readable text and formal, machine-readable axioms [9]. An ontology expresses, for a particular domain, the set of terms, entities, objects, classes and the relationships between them, and provides formal definitions and axioms that constrain the interpretation of these terms. By making explicit the implicit definitions and relations of classes, objects, and entities, ontology engineering contributes to knowledge sharing and re-use. [9-10]. An ontology permits a rich variety of structural and nonstructural relationships, such as generalization, inheritance, aggregation, and instantiation [11]. Shared ontological commitments by users and systems guarantee consistency. For example, an XML schema expresses a machine-readable ontological commitment. Computing services that use the same XML schema consistently exchange information and have some degree of inter-operability. However, what information XML elements represent is not specified, and relationships between elements are limited to

enumeration and nesting. Errors due to ambiguity in what elements actually represent may occur and there is a lack of flexibility in representing relationships.

Ontologies have been represented in machine-readable, frame-based and description logic languages, including Knowledge Interchange Format (KIF), Resource Description Framework (RDF), DARPA Mark-up Language + Ontology Inference Language (DAML+OIL), Web Ontology Language (OWL), and others [12]. It is the declaration of a classification system with classes, sub-classes, taxonomies, definitions, properties, relationships and axioms that taken together specify a particular ontology.

The set of classes and terms presented here form some steps towards defining an ontology of scientific information that is becoming increasingly needed in grid environments. A large part of this information is contained in metadata schemas that describe the data and are often, but not always, found with it. New schemas are developed and legacy schemas are used with large costs associated with integration.

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. For instance, information about the instruments and experiments that produced the raw data and what transformations it was subjected to in the course of its life may be useful, along with information regarding the project for which it was produced, and who the principal investigator of the project was. Information about a user of the data is also necessary to provide users' credentials and access rights to particular data sets.

Our ontology is intended to provide a basis for classifying and retrieving data files, collections and information about the files and collections based on content for use in a grid context. By declaring the meaning of metadata terms and how they are related using underlying abstract concepts, we try to remove some ambiguity in the choice for what data is to be analyzed, what is contained in a file or a collection of files, and what metadata may be searched. By facilitating the decision-making process before data files are transferred to their point of service, this ontology aims to save time and computing resources, and bring more transparency to a scientific user.

Grid architectures are service-oriented and emphasize operations that can be performed on data using the associated metadata schemas, rather than focusing upon the content of metadata schemas and relationships between schema elements. Much metadata in grid architectures is implicit, often contained within each service, and may be implemented in XML schemas. Documents describing schemas for a particular system may exist but not always. Metadata schemas are found in database tables and storage systems' back ends 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. Redundancy, overlap, and gaps may occur without the explicit knowledge of the user, leading to interpretation errors. By expressing relationships between metadata elements this ontology attempts to remove some ambiguity.

3. Declaration of entities and relationships

| | |
|----------------|--|
| Pedigree | Represents a line of ancestry from creation through various transformations to arrive at the current data set. It also includes information related to the scientific project and data identity. |
| Scientific_Use | Describes how a scientist used the data, what experiments were performed, what were the parameters and configuration of models. |
| Dataset | Describes data typically stored storage facilities, and may include parameters, location, and the study that produces this data. |
| Service | Concerns how a service may be invoked and what its capabilities are in a gridded architecture. |
| Access | Concerns whom is allowed to access the data, security and authentication. |
| Other | Includes annotations, comments, and evaluations. |

Table 1. Classes

Table 1 presents a set of abstract classes representing concepts that reflect best practice for scientific information in a grid setting. These high-level classes

pertain to information that is common and may be required in many different domain areas. The instantiation and specialization of these classes will depend on domain areas and applications. For instance, Pedigree, Access, and Service may have the same specifications across several disciplines, but Scientific Use and Dataset may be unique. Finer grain ontologies focus on domain-specific concepts, and are designed by domain experts. Domain-specific ontologies are mapped to the above abstractions using relationships such as:

- Dataset is_associated_to Pedigree
- Dataset uses Service
- Scientific Use is_associated_to Pedigree

The classes in Table 1 are supported by the service-oriented focus of grid middleware and may be mapped to tools found in the Globus toolkit. The ontology is intended to be architecture independent. The service entity may contain information for a grid service to be invoked in a grid setting and may be based on the Open Grid Services Infrastructure to enable inter-operability. (OGSI) [13]. Ontology services are based on content and designed to help community users find and query services that advertise their capabilities. More precisely, OGSI allows connection to an already known service, and ontology services allow discovery of that service. While the specifications found in [13] mostly concern access protocols, ontology services enable access to content for browsing, discovering entities, phrasing queries, and for reconciling new or customized metadata schemas with minimal manual input. Other concepts apply to data and applications that may be accessed by the service. Scientific use and dataset concepts tend to be domain-dependent whereas pedigree and access may be independent.

Figure 1 presents classes designed for the ESG project. ESG uses the abstractions presented in Table 1 to suit project requirements. Some mappings between ESG objects and abstractions are presented here. ESG entities that are not represented in Figure 1 include Activity and Format. Relationships between ESG classes include:

- is_associated_with (transitive and symmetric),
- is_a_set_of,
- has_parent, Inverse_of_has_parent,
- has_sibling,
- has_role,

- has_parameters.

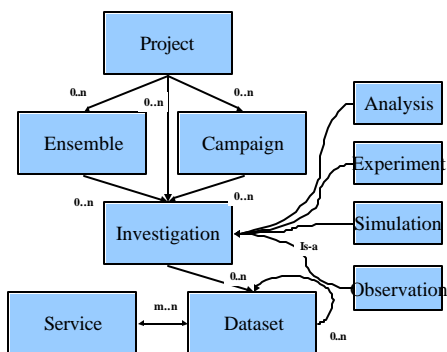


Figure 1: Main ESG classes

Concepts related to pedigree, scientific use, datasets, and access are discussed. The ESG ontology has been developed using the OilEd editor [14] to produce a DAML+OIL ontology. Protégé-2000 [15] was also used. As a domain specific ontology, the ESG ontology contains entities that are common to most collaborative projects, such as access, pedigree, and domain-specific concepts such as datasets and scientific use.

3.1 Pedigree

Pedigree describes the origin, identification, and history of a data collection and individual files. Pedigree may provide “a line of ancestors,” identity, provenance, and project information. In the case of data sets assembled in data collections, pedigree may also include information about the building of this collection. Identity is currently best described with the Dublin Core (DC) Metadata Element Set [16] elements as the use of DC is slowly spreading through scientific communities. When associated with scientific use a pedigree may allow derivations of a data set based on earlier data sets recorded in pedigree information. The pedigree trace facilitates re-constructing the workflow of an experiment by following the transformations of data through a series of related experiments (input to a model under certain conditions, output of an experiment, perturbing an input condition). Pedigree may be a quality control mechanism for a creator of data files and collections who wants to review forgotten details about how s/he arrived at certain results.

| | |
|------------|---|
| Identity | DC:Name or URI DC:Creator DC:Contributor DC:Publisher DC:Time information DC:Version information |
| Provenance | Other datasets used as input for this data or from which this data was derived. Attributes used for collection building. |
| Project | A project is an organized set of investigations that produce data. The scope and duration may vary from a few month to multiple years. It typically has one or more principal investigator and a source of funding. |

Table 2. Pedigree

Project and Person represent ESG’s concepts for identity. DC elements provided in the OilEd editor are used (title, creator, subject, description, date, version). Other useful DC elements may be Source, Contributor, and Resource Identifier. A Person is restricted by the relationship has_roles such as Principal Investigator, Data Manager (someone who manages data-intensive storage facilities and/or content used in the Earth Sciences). Person is also restricted by the relationships:

- Person is associated_with Dataset, Investigation, Project.
- Project is associated_with Investigation.
- Person is intended to facilitate a quick access to datasets already known to be associated with a person, in particular a scientist’s own past datasets.

In ESG the class Provenance is not used as such but Provenance is represented in subsets of classes found in scientific use such as Simulation, Parameters, and instances of Datasets. The Collaboratory for Multi-Scale Chemical Science (CMCS) [17] (not discussed in this paper) implements the concept of pedigree with the CMCS Explorer, a pedigree browser.

3.2 Scientific Use

This is the category of most interest to scientists. Scientific use is domain-specific information necessary for the scientific analysis of data sets. It varies with disciplines and includes variables, parameters, run-time

conditions, data formats, and other characteristics. In particular, elements in existing metadata schemas need to be reflected. Queries for obtaining datasets are often based in part on domain specific attributes. In ESG, scientific use is represented with parameters that include model configuration, input datasets, initial and boundary conditions, time period, spatial coverage, measurement ranges and units. ESG scientific use applies to Investigation, Campaign, Ensemble, and Parameter. An Investigation is defined as an activity that produces data within a project. Simulation, Observation, Experiment, and Analysis are sub-classes of Investigation and inherit the restrictions that apply to investigations in addition to some restrictions related to Parameters. (Tables 3 and 4). The relationships `has_parent` and `has_sibling` apply to simulation.

| Investigation | |
|--------------------|--------------------|
| Is_associated-with | Datasets Person |
| Has_parameter | Parameters |

Table 3. Scientific Use

Table 4 describes sub-classes of investigation and the parameters used to distinguish them.

| Sub-classes of Investigation and restrictions | |
|---|---|
| Simulation | Model configuration. Input datasets. Initial and boundary conditions. |
| Experiment | |
| Observation | Measurements |
| Analysis | Input Datasets Processing History |

Table 4. Investigation

In addition, the ESG class Ensemble is defined as a set of closely related simulations where aspects of the model configuration are held constant while the initial or boundary conditions vary out of a normal range. A Campaign is a set of observational activities that share a common goal (e.g., observation of the ozone layer during the winter/spring months), and are related either geographically (e.g., a campaign at the South Pole) and/or temporally (e.g., measurements of rainfall at several observation stations during December 2002). In ESG, the following relationships apply:

- Ensemble is_a_set_of Simulation,
- Campaign is_a_set_of Observation,
- Simulation has_parent Simulation,

- Simulation has_sibling Simulation.

3.3 Dataset

Scientific projects are associated with datasets, e.g. data containers that are the outputs and inputs of experiments and investigations, and the outputs of observations as raw data. Datasets can be composed in collections, in aggregation, and slicing where only a subset of a data file is of interest based on certain variables. In a grid environment, datasets are often distributed between several storage facilities, may be duplicated for performance reason or to avoid a single point of failure, and for some grid projects including ESG may be measured in terabytes. Datasets may be represented in various formats including simple ASCII or binary, or more advanced, network transparent self-describing forms. Domain-specific conventions may be adopted as well. There are often multiple domain-specific formats, some of them expressed in XML schemas. Datasets have locations on different types of storage systems. The following relationships apply in ESG:

- Investigation is_associated_with Dataset.
- Person is_associated_with Dataset.
- Dataset has_parameters Parameter.

The `has_parameter` relationship is implemented for data discovery so that a scientist may view only datasets that contain a particular parameter (cloud, latitude are examples of parameters). A dataset may represent a single file or a collection of files compiled by a user.

3.4. Access

The concept of access describes the information necessary to provide secure, fine-grained access to restricted resources for a user. It also describes the process to gain access and perform some operations on that resource. In grid architectures access may be supported by the concepts of community and access policies [18]. Data and computing resources reside at different institutions, but each institution may have different access policies for its own resources. Communities are sets of people and/or institutions sharing rules that define access and use of resources within a distributed environment [1]. For individual users, the community they belong to determines access to a resource. An institution may in turn grant access to

the entire community without managing each individual user's access. Communities may also be divided in groups. User authentication is based on certificates.

| | |
|-------------------------|---|
| Community | The community of users and processes engaged in a collaboration that necessitates sharing distributed resources. |
| Policy | Statements defining the community policy or policies regarding members' access. |
| Resources | A list of resources that members of the community may access. |
| Privileges | A list of privileges assigned to members or group of members |
| Member list | A list of members in the community. |
| Group | The group within the community to which the member has been assigned if groups exist for this community |
| Certificate Authority | The designated person to notify when a new member needs a certificate. |
| Community Administrator | The designated member of the community who administers the member list and assigns privileges to individual members and groups. |

Table 5. Access

4. Discussion

At the time of this writing a final version of an ESG metadata schema implemented in XML is nearly completed. An object model and a relational database backend that contains instances of schema elements will also be available. The proposed ontology above (and all ontologies) expresses relationships between ESG metadata objects and constrains their use. Relationships are themselves objects, so that their properties can be restricted and their usage constrained: objects and relationships are formalized and can represent any concept useful for describing a domain. Given the formalization of objects and relationships, object definitions are machine-readable. Protégé-2000 and OilEd were both used. Protege was preferred for its graphical representation capabilities with the

OntoViz plug-in. The lightness of OilEd appeared more desirable in a first iteration of the ontology. The more robust and extensive capabilities of Protégé may be used in the future.

Ontology development tools also allow developers to perform validation of schemas against the proposed ontology. This verifies that object definitions in a metadata schema do not conflict and may highlight implicit definitions to be resolved during further iterations. Reasoning on definitions also becomes possible so that new and existing schemas can appropriately co-exist and be related in applications without errors on what an element represents. For this reason, the ontology also enables the sharing and re-use of tools developed separately, as is planned for ESG and NERC. Reconciliation between the NDG-CLRC and ESG schemas may use the proposed ontology.

The ESG architecture is currently build on the Globus Toolkit 2.2 [5]. Access mechanisms are provided through the Community Authentication Service (CAS) [19], and ESG concepts regarding access are directly related to CAS. Minimal pedigree information is provided by the Metadata Catalog Service (MCS) [20], some of which is encoded in the Dublin Core. MCS has also implemented some concepts for scientific use and datasets such as user attributes that contain some dataset parameters. NetCDF [21] is the first data format for climate model data implemented in ESG. The NetCDF Mark-up Language (NcML), an XML schema for NetCDF, has been developed. [22]. Table 6 summarizes some ESG services and schemas.

| Ontology Entities | ESG services | Schemas used |
|-------------------|--------------|--|
| Pedigree | MCS | Dublin Core Unqualified Elements Set. Other. |
| Scientific Use | | NetCDF variables. Data Format: NetCML |
| Access | CAS | |

Table 6: Summary of ESG services

ESG services are quite different from the grid services described in the Open Grid Service Infrastructure (OGSI). For OGSI, a grid service is a set of

conventions for the characteristics of a particular service, its capabilities, and lifetime management of the service in an open grid environment [13]. ESG services are categorized as metadata services, transport services, access services, and application level services in the high-level view of the ESG architecture. In addition, in the ESG object model (Figure 1) a service links datasets to the protocols and storage facilities that can be used to obtain a dataset of interest.

5. Conclusion

This paper presented an ontology for scientific information in Grid Computing. The concepts of pedigree, scientific use, dataset and access were designed and examples were given for the Earth System Grid. The examples illustrated how a specific grid project used and modified ontological concepts for its domain area. While pedigree information identifies other data files or data sets, scientific use described the conditions of production. Other domain-specific information characterizes datasets. Future tasks in the ESG project include validation and consistency checking of schemas to the ontology, and reconciliation with NERC-CLRC.

Research tools and grid systems are being developed for scientific laboratories in application domains such as climate and earth sciences, physics, chemical science and others. In addition to ESG, other laboratories use concepts found in the present ontology such as pedigree in CMCS [17]. GriPHYN, the Grid Physics Network has implemented some measure of provenance in the Chimera system [23]. Formal specifications are being designed for data and resource sharing as is envisioned in the Semantic Grid, and the European Data Grid. Lower levels of ontological granularity for domain-specific schema representation are also needed.

References

[1] The Earth System Grid. <http://www.earthsystemgrid.org/>.
 [2] 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
 [3] The Grid: Blueprint for a new Computing Infrastructure. Ian Foster and Carl Kesselman, eds. Morgan Kaufmann, 1999.

[4] Distributed Oceanographic Data System, <http://www.unidata.ucar.edu/packages/dods/>.
 [5] The Globus project. <http://www.globus.org/>.
 [6] The Earth System Grid: Turning Climate Datasets into Community Resources. AMS 2002.
 [7] The NERC Data Grid. <http://ndg.badc.rl.ac.uk/>
 [8] The CLRC e-Science Center. <http://www.e-science.clrc.ac.uk/>
 [9] Gruber, T., "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisitions* 5, (May 1993): 199-220.
 [10] A. Gomez-Perez, 1998. "Knowledge Sharing and Re-Use," In *The Handbook of Applied Expert Systems*, J. Liebowitz, ed. Boca Raton, (1998), pp.10:1-10:36.
 [11] M.N. Huhns and M. P. Singh. "Ontologies for Agents," *IEEE-Internet Computing* 1, no. 6 (Nov.-Dec. 1997) pp. 81~83.
 [12] KIF, <http://logic.stanford.edu/kif/kif.html>, RDF, <http://www.w3.org/RDF/>, DAML+OIL, <http://www.daml.org/>, OWL, <http://www.w3.org/TR/2002/WD-owl-features-20020729/>.
 [13] S. Tuecke, K. Czajkowski, I. Foster, J. Frey, S. Graham and C. Kesselman, D. Snelling, P. Vanderbilt, "The Open Grid Service Infrastructure, (OGSI)" (Draft, February 17, 2003), <http://www.globus.org/ogsa/>.
 [14] The OilEd Editor, <http://oiled.man.ac.uk/>.
 [15] Protégé-2000, <http://protege.stanford.edu/>.
 [16] The Dublin Core Metadata Element Set V1.1 (DCMES). <http://dublincore.org/usage/terms/dc/current-elements/>.
 [17] Collaboratory for Multi-Scale Chemical Science. <http://cmcs.ca.sandia.gov/index.php>.
 [18] L. Pearlman, V. Welch, I. Foster, C. Kesselman, S. Tuecke, "A Community Authorization Service for Group Collaboration," *IEEE Workshop on Policies for Distributed Systems*, 2002.
 [19] The Community Authorization Service, <http://www-fp.globus.org/security/CAS/>.
 [20] A. Chervenak, E. Deelman, C. Kesselman, A. Pearlman G. Singh, "A Metadata Catalog Service for Data Intensive Applications," Version 1.0, June 26, 2002 Draft.
 [21] NetCDF, Network Common Data Form, <http://www.unidata.ucar.edu/packages/netcdf/>.
 [22] The NetCDF Mark-up Language. <http://www.vets.ucar.edu/luca/netcdf/index.html>.
 [23] GriPHYN: Grid Physics Network, <http://www.griphyn.org/index.php>.

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.