

# Data Grid Discovery and Semantic Web Technologies for the Earth Sciences

## LINE POUCHARD

*Computer Science and Mathematics, Oak Ridge National Laboratory, P.O. Box 2008,  
Oak Ridge, TN 37831-6367. USA.*

[pouchardlc@ornl.gov](mailto:pouchardlc@ornl.gov)

Telephone: +1 865-574-6125

Fax: +1 865-574-0680

## ANDREW WOOLF

*Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire, OX11 0QX. UK.*

[a.woolf@rl.ac.uk](mailto:a.woolf@rl.ac.uk)

Telephone: +44-1235-448-027

Fax: +44 1235-445-808

## DAVID BERNHOLDT

*Computer Science and Mathematics, Oak Ridge National Laboratory, P.O. Box 2008,  
Oak Ridge, TN 37831-6016. USA.*

[berhno1dtd@ornl.gov](mailto:berhno1dtd@ornl.gov)

Telephone: +1 865-574-3147

## Acknowledgements.

The submitted manuscript has been authored by a contractor of the U.S. Government under Contract No. DE-AC05-00OR22725. Accordingly, the U.S. Government retains a non-exclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. The NERC Data Grid is funded under the UK e-Science program through a grant (NER/T/S/2002/00091) from the Natural Environment Research Council.

The technologies powering the Earth System Grid and the British NERC Data Grid were developed by many members of each team. In particular, the authors would like to thank: Shishir Bharathi, David Bernholdt, David Brown, Kasidit Chanchio, Meili Chen, Ann Chervenak, Luca Cinquini, Bob Drach, Ian Foster, Peter Fox, Jose Garcia, Carl Kesselman, Don Middleton, Veronika Nefedova, Arie Shoshani, Alex Sim, Gary Strand, Dean Williams, working under the guidance and sponsorship of Mary Ann Scott in the US, and Kerstin Kleese van Dam, Bryan Lawrence, Roy Lowry, Ray Cramer, Marta Gutierrez, Siva Kondapalli, Susan Latham and Kevin O'Neill for the British team.

Abstract

This paper describes scientific data discovery for the earth sciences in the context of data grids and grid computing. Requirements and use cases illustrate current challenges due to size, distribution and minimal annotation of data. Semantics and the characterization of provenance in large data archives are discussed. The targeted community of users is also discussed. Solutions implemented by the Earth System Grid and the National Environment Research Council Data Grid include a prototype ontology, metadata schemas, search mechanisms and discovery architectures. The use of Semantic Web technologies has facilitated the development of meaningful annotations of data content, and opened the door to data discovery in federated systems.

### ***Keywords***

*Semantic Web. Semantic Grid. Grid Services. Ontologies. Data Discovery. Earth Sciences. Scientific Information Retrieval. Standards.*

### **Abbreviations**

CCLRC, (UK) Council for the Central Laboratory of the Research Councils; ESG (US) Earth System Grid [ref 1]; FGDC, (US) Federal Geographic Data Committee [ref2]; ISO, International Standards Organization; NASA, (US) National Aeronautics and Space Administration; NERC (UK) National Environment Research Council; NDG, (UK) NERC Data Grid; OWL, Web Ontology Language; RDF, Resource Description Framework; XML, Extensible Mark-up Language; WSDL, Web Service Definition Language; OWL, RDF, WSDL, XML are W3C candidate recommendations or standards [ref 3].

## **1. Introduction and background**

In emerging grids and Grid Computing [ref4], shared, distributed and heterogeneous computing and data resources 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 [ref 5]. The management of these resources is complex, time-consuming, and not subject to centralized control. In data-intensive scientific domains, such as the earth sciences, high-energy physics, and astronomy, many terabytes of data are being acquired from simulations performed on supercomputers and from experiments/observations 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.

## Figure 1

The Earth System Grid is developing a virtual collaborative environment based on Grid technologies to facilitate analyzing the impacts of global climate change at national laboratories, universities and other laboratories (Figure 1). ESG is a project of the U.S. Department of Energy Scientific Discovery through Advanced Computing program. ESG provides access to data produced by earth and climate science simulations through a Web portal. Climate scientists and researchers utilize distributed resources to discover, access, select, and analyze model data produced by simulation runs and stored in large archives. Semantic Web technologies may prove useful for smarter and more flexible tools to address the real life challenges encountered in scientific data management.

ESG is also pursuing collaboration with the British NERC Data Grid (NDG) and CCLRC [ref 6]. CCLRC's mandate include several sciences. The NDG project is motivated by a broadly similar aim to ESG – the need for end-user scientists to seamlessly search for and access a wide variety of earth sciences related data. In both ESG and NDG discovery and access extend across multiple geographic locations and administrative domains, including the National Center for Atmospheric Research, and the Department of Energy laboratories in the US, and the NERC “Designated Data Centres” in the UK. Some requirements and architecture design are also similar. CCLRC and ESG plan to leverage tools from each other.

This paper presents challenges for searching and retrieval of scientific information and solutions implemented by ESG and NDG using metadata services. Challenges posed by the provenance of datasets and federation of services between projects increase in distributed data grids with choice of resources (storage sites, catalogs, and servers), sizes and multiplicity of users and are addressed. Scientific users' needs to locate files prior to

downloading, based on content, stored in geographically distributed archives, provenance, and the role of ontologies in scientific grids are of particular relevance to the Semantic Web and discussed here. The paper also discusses a prototype ESG ontology, metadata services and schemas, the ESG computer architecture, and proposes directions for incorporating more semantics. Security, data transfer, and web portal design are not within the scope of this paper.

## **2. Data Discovery**

### **2.1. Requirements**

User requirements were established by close collaboration between computer scientists and domain experts. Tracing provenance [ref 7], a concept that loosely describes where a file comes from and what transformations it went through, becomes crucial. It may include names and versions of simulation models, resources used in production, computers where models are run, and/or names of funding agencies. Searches are expected to point to datasets based on search criteria such as date and time coverage, presence of variables, type of simulation models, creators of datasets, and related datasets. Access through a single point of entry from a scientist's desktop is required.

ESG users are climate scientists at national laboratories, other government agencies, and universities around the country and abroad, including climate science, oceanography, land surface, sun-earth interactions, and other disciplines included in the NASA Global Change Master Directory [ref 8]. Earth scientists providing expertise for the Inter-governmental Panel on Climate Change, the architects of the Kyoto treaty, are a main target user group. A motivation for the development of ESG is to improve access to online resources and community data for users who lack of awareness of what is available. 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 must also be seamlessly initiated from a desktop machine, often through a site other than the user's home site and the location where the simulation was run. This occurs when users have access to a variety of remote supercomputing resources on which they perform analysis. Because of the size of datasets and therefore the length of the transfer, 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 may run for one to several weeks and consume many computational resources and man/hours.

NDG Data Providers maintain data holdings under separate administrative and policy domains. These include, for instance, the holdings curated by the British Atmospheric Data Centre and British Oceanographic Data Centre. Users require the ability to be able to search across a number of conceptual dimensions associated with earth science data (Figure 2); for example rainfall measurements from ground-based weather radar (data production tool); or for ozone mass-mixing ratio from the ERA40 reanalysis (data activity), or for current mooring data (observation station type) in the Pentland Firth. Filtering on physical parameter and location (temporal and spatial) is also required.

## **Figure 2.**

ESG data is binary data only obtained by running climate simulation models (processed data). ESG is not currently expected to manipulate raw data from observation stations. Processed data is used to create new models with the effect that at the end of a chain, it may be difficult to determine which analyses a dataset went through. Some datasets are linked to each other by model configurations, parameter variations, (geographic and atmospheric) grids and some datasets are part of collections or ensembles. Current practice in this area depends heavily on the involvement of particular individuals (calculation managers) for the discovery of available data. Data sizes already are barely manageable and data loss will occur if discovery mechanisms are not soon and greatly improved, i.e. the data exists somewhere but cannot be found. As of July 2003, the estimated total volume of data to be created by running the necessary simulations for next

round of the Inter-governmental Panel on Climate Change studies is 18.91 terabytes corresponding to 3230 model years distributed over three storage sites as follows:

- National Center for Atmospheric Research, 530 model years, 1530y, 8.961 terabytes,
- Lawrence Berkeley National Laboratory, 600 model years, 600y, 3.514 terabytes,
- Oak Ridge National Laboratory; 1100 model years, 1100y, 6.443 terabytes.

As well as simulated data produced by state-of-the-art numerical models, NDG needs to facilitate access to observational data (including remote-sensing imagery). These data are often complex, irregular, and have rich and important semantics of their own (e.g. a single marine Conductivity-Temperature-Depth profile measuring temperature and salinity of seawater with depth) may be part of a longer hydrographic section, and the data may be combined with that collected from another ‘underway’ instrument to build a detailed picture of synoptic upper-ocean temperature structure). The richness of data types being supported in NDG indicates the need for a data model with semantics. Requirements capture indicated also the need for individual research groups to be able to share their data by registering into the NDG infrastructure. A number of datasets (and metadata in some cases) have restricted access, and so security is a fundamental concern.

## **2.2. Search and Retrieval Use Cases**

The search and retrieval of datasets generated by earth science simulations and observations are a primary functionality of ESG and NDG. The ability to locate and obtain datasets as easily and seamlessly as possible is crucial to climate and other scientists who handle large files. The time currently needed for locating a file must be shortened and the human input automated. Search and retrieval are based on metadata schemas. Fine granularity in the representation of users and actions was essential for the usability of schemas.

Five information retrieval scenarios were designed for ESG.

- (1) A user browses dataset catalogs and wants to know details related to simulation model configuration, variables contained, and years of coverage, for some datasets without downloading them.
- (2) A computer application creates the necessary metadata as datasets are produced. The application creates the new metadata file uniquely identifying the data according to metadata.
- (3) A data manager searches for datasets he registered and stored last year.
- (4) A scientist wants to re-visit datasets she grouped in one view .

More complex searches are envisioned. These include:

- (1) identifying datasets containing a given variable across datasets with unrelated schemas.
- (2) returning slices of data for files containing the variables “wind” and temperature” at particular 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) returning the datasets above from data archives held in repositories. 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.

In practice, requirements for ESG metadata services include:

- model run descriptions (including input scenarios and input data), model configuration information, and model components (atmosphere, ice, ocean),
- input datasets
- pointers to documentation,
- sites where simulation take place the models are run,
- and people who carried out the model integration and submission to archives.

The ability to capture relationships that link datasets in ways such as “parent,” “child,” and “sibling” was also important.

Requirements analysis indicated the need for NDG to provide search facilities across standards-based metadata schema (such as the Dublin Core [ref 9], FGDC, ISO 19115, and the GEO profile of the Z39.50 protocol). Thus mappings are being made from the NDG schema, and these standard formats will be supported through a metadata export interface.

### **3. The Need for Semantics**

For information discovery the Semantic Web may serve 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). Another classic example is when this user searches travel information and reservations in a search powered by composeable Web services based on semantics. This need places her in another community (travelers). By contrast, scientific communities tend to be relatively small (i.e. not all the people looking for travel arrangements) and narrowly defined by domain expertise when compared to the Web communities above. They exist prior to and independently from a request for information, and are much more persistent. The emerging challenges of science require team efforts, inter-disciplinary collaborations between geographically disperse groups, and sharing limited resources such as supercomputers and large instruments. Integrated computer applications and single point of entry through multi-purpose portals accessed from a desktop are also crucial. Although this community may be more precisely defined their needs are more daunting.

Metadata for scientific information is any information scientists may need for making decisions about analysis, studying the production of data, detailing actions and results in publications. Here, metadata refers to the list of objects and the object schema that contains all available description items for a given data. A metadata instance refers to the schema item used for a particular dataset and its associated value. For instance, “simulation name” is a metadata item, “Parallel Climate Model B04” is the metadata

instance for dataset X. Provenance information may be represented in one or several metadata items.

Provenance of a dataset is known in an ad hoc fashion sometimes held in a scientist's and/or the archive administrator's head. This 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. Information about the data such as its time periods, versions, and variables may be stored with binary data, so that the only access is by transfer and examining file content. Lists of datasets may be contained in electronic catalogs with little known information beyond the filename. Scientists typically know what to expect from a simulation model and trust known simulation data producers. They rely on memory and publications for the characteristics of the data. However, this method is no longer practical and reliable due to the size and multiplication of simulation datasets produced on the newest supercomputers.

From a data perspective, Grid tools such as the service-oriented Globus Toolkit [ref 10] have emphasized operations such as high-speed and secure transfer to and from distributed mass storage. Metadata for grid data is often implicit, and sometimes used within a grid service, but not described. This renders effective collaboration and data sharing difficult. Some metadata schemas are found in database tables and storage systems that are not usually directly accessible to a scientific user and may be of limited use for discovery purposes. This state of things makes metadata difficult to access and compare. Such metadata contains little semantics beyond an entity-relationship model. At best, metadata is described and available in XML with a data dictionary. Redundancy, overlap, and gaps may occur without the user being aware of it, leading to interpretation errors. By expressing relationships between metadata elements and increasing interoperability between earth science metadata, ontologies attempt to remove some ambiguity. Adding support to search mechanisms, content descriptions and all

annotations that help characterize the data and computing resources contained in metadata are becoming a major focus of service frameworks such as OGSADAI [ref 11].

A number of markup languages are being developed to enable the description of data file contents. They are all primarily syntactic in nature. The Earth Science Markup Language (ESML) [ref 23] provides a mechanism for describing the structural contents of various earth science file formats (GRIB and HDF-EOS, as well as ASCII and binary). ESML software libraries use such a description to enable access to the file's data through a single API. In a similar manner, the netCDF Markup Language (NcML) [ref 26] describes the contents of netCDF files. The Data Format Description Language (DFDL) project [ref 27] is an ambitious attempt to develop a general-purpose file description language. Semantic descriptions may be layered on top of any of these file format description languages to facilitate the transition from data to information. For example, semantic enhancements to ESML [ref 24] will enable file contents to be identified with terms from a domain ontology (e.g. "latitude" or "time"). Semantically meaningful operations (such as 'subsetting') may then be performed automatically. Recent GIS developments [ref 25] start from a position of defining *a-priori* important conceptual data types (called "feature types"). Higher-level semantic services (e.g. coordinate transformations) may then be invoked, and chained together. To apply this approach to earth science data requires a mechanism for connecting legacy data files to semantic feature instances. This is the approach adopted by NDG [ref 21].

## 4. Results

### 4.1. ESG and NDG semantics in practice

#### Figure 3.

A prototype for an ESG ontology [ref. 12] was developed using Protégé-2000 [ref 13]. It specifies broad categories for content information found in ESG and other Grid projects. The ESG ontology contains the disjoint concepts of Pedigree, Scientific Investigation, Datasets, Service, Access, and Other (figure 3). The ESG Pedigree represents identity and line of ancestry (provenance) for other entities in the ontology. Provenance may

apply to a dataset or an investigation. Using pedigree relationships, people and institutions are associated with scientific investigations by roles such as PI or funding agency, and with datasets by roles such as data manager or data publisher. Provenance is a subclass of pedigree and records names or IDs of datasets that served as input or output for a particular simulation. Some pedigree information uses the Dublin Core. A Scientific Investigation describes an activity that produces data such as a simulation, an experiment, an observation, or analysis and specifies all information that is pertinent to data production. As ESG focuses on simulations, simulation slots in ESG describe model configuration, input datasets, initial and boundary conditions and sites and machines where the simulation was run. Dataset describes a container for data that may correspond to a single data file, a collection of related data files, or a set of entries in a database. ESG datasets have a format, temporal and spatial coverage, a simulation calendar, and parameters.

## Figure 4

Two main relationships in the ESG ontology include (Figure 4):

- `isPartOf`: a dataset is part of an investigation.
- `generatedBy`.: Dataset L is generated by Dataset P.

Thanks to Provenance and Scientific Investigation information, a user may trace the conditions under which a particular dataset has been produced, including simulation input datasets, simulation models, or the information associated to a data producer (Figure 5). Provenance and Scientific Investigation 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 but this information is not organized, recorded, and directly accessible with datasets. Provenance information may also be used for verification of his own data and models by a scientist instead of performing frustrating searches in old notes.

## Figure 5

In the ESG prototype ontology (Figure 3), a service is a coherent functional capability that may be realized for example through an API or a Web service. It associates earth science data formats with servers capable of delivering data in that format or processing including operations as subsetting in coordinate space, visualization, and evaluating expressions. A service may be provided by several servers, and a single server may be able to treat several types of formats. The Access entity of the ontology describes which person, group, and computer application is allowed to access ESG data using security and authentication information. The Other entity represents (mostly) manual annotations, notes and references.

The logical separation of ontology entities between domain-specific metadata for ESG and what may be used in other Grid projects has been a leading principle in building the ESG prototype ontology (Figure 3). Scientific Investigation and Provenance may be domain-specific whereas specifying Access, Dataset, and Pedigree may be common to several grid projects. For instance, while sub-classes of Scientific Investigation, such as Experiment, and Observation may apply to other Grid projects, Campaign and Ensemble may not. Dataset metadata such as associated project, and dataset owner may be common, but not parameter metadata. As tools suitable to several projects such as metadata catalog services and replica location services (and their later incarnations in the Globus toolkit and OGSADAI) become more common, metadata schemas used by these tools may be re-used to build ontologies in other grid projects. For instance, metadata items suitable to describe “logistics” or “house cleaning tasks” may be re-usable.

For meaningful data use, NDG has constructed a data model incorporating the following core data semantics: “structure” (through nested hierarchies of multidimensional arrays), location in time and space, and storage descriptors (to enable encapsulation of file formats, storage location etc). Data access services leverage this data model to virtualize data resources – a key pattern of Grid computing. A simple example of such virtualization is the ability to aggregate component model data files (e. g. along a spatial or temporal axis) into a larger logical array. It also hides file format details, so that data

stored in NASA-Ames files, for instance, may be exposed in whatever format the User requires.

## 4.2. ESG schema

### Figure 6

ESG developed its own XML metadata schema focusing on earth sciences modeled data (See Figure 6). ESG evaluated several existing data description solutions for use with earth sciences data and found the following:

- The Dublin Core was not rich enough to support scientific data because its primary purpose is to describe papers and electronic publications (such as web pages). It has been used for parts of the pedigree information in other Grid projects.
- The ISO standards of the FGDC proved too detailed for ESG purposes and timely implementation. There are 339 elements in the schema, with, for example, 12 concepts alone for characterizing versions of the metadata file [ref 14 ].
- The Data Interchange Format (DIF) was the closest to ESG needs, and future mappings between ESG and DIF are expected based on user requests. The DIF controlled vocabulary focuses on representing experimental and observational data, but model data is not well represented. In particular, model configuration and model run time information are absent. Parent relationships for characterizing ensemble runs exist but do not permit sibling datasets.

## 4.3. Collections and file names in the ESG schema

The ESG schema focuses on describing collections and search and discovery of collections. A collection may be formed by files, datasets and/or other collections. A collection is also a dataset described by its own metadata instances. A group of datasets each described with its unique metadata instances may be assembled in a collection. Collections and the selection criteria vary. Criteria for building collections include relations between files such as parent, child, and sibling relationships, all of which are allowed in the ESG schema. Siblings are datasets with a common parent. Other relations between files that are of interest to a user or collection builder may also constitute criteria

for inclusion. For instance, collections may be based on multi-dimensional coordinates, time-related coverage, or ensembles of model runs.

## Figure 7

ESG uses logical file names to reference datasets and physical file names to locate them. In ESG, file names may already indicate the name of the model, and type of model (e.g. atmosphere), and the dataset format, but this content description contained in file names is limited. A query to the ESG discovery services returns logical file names according to search criteria. The logical file may represent a single file, a set of logically related files, or a dataset, such as a collection. The logical file name of interest points to a set of physical files, possibly in different archives. The user then chooses a location from where to download the file or collection (Figure 7).

### 4.4. High level architecture

The ESG data discovery and transfer is based on the Open Grid Service Architecture [ref 15]. It is component-based with components distributed at various ESG sites communicating through Simple Object Access Protocol (SOAP) for searches and metadata requests. Data transfer and download use GridFTP [ref 16]. Figure 8 presents a high-level view of the discovery architecture in ESG.

## Figure 8

A user submits a search to the ESG portal that transmits it to the underlying discovery service located at National Center for Atmospheric Research. This service parses the query to be sent the Metadata Catalog Service that returns zero or several logical file names to the Discovery Service. Metadata associated to each logical name is also returned to the portal. Logical file names are sent to the Replica Location Service that returns to the portal physical file names and a Universal Resource Locator (URL) for the files corresponding to the logical file name. The user may chose to download some files. At the time of this writing the Metadata Catalog Service is being migrated to the Open

Grid Services Architecture Data Access and Integration (OGSADAI), a Grid service standard from the Global Grid Forum.

#### **4.5. NDG solutions**

NDG Data Providers maintain detailed metadata for their datasets, compliant with an NDG metadata schema [ref 20]. Tools will be developed to facilitate metadata creation and management. Discovery-level metadata (Dublin Core and FGDC DIF) is generated by Data Providers as a summary transformation from the detailed metadata, and harvested by one or more Discovery Services. In this way, individual data catalogues may be federated, and searched centrally. The protocols of the Open Archives Initiative [ref 17] are used for metadata harvesting. An NDG authenticated user may then search discovery metadata and browse detailed metadata, and browse or download data to which they have access. Data delivery services use the NDG data model [ref 21] as a means of encapsulating storage details.

The NDG metadata schema includes the following high-level entities: data activity (e.g. funded project, field campaign, reference simulation), data production tool (e.g. instrument, model), and observation station type (land station, mooring, aircraft, ship, etc). Relationships between these core entities enable searching to be carried out across these dimensions. A taxonomy of metadata in NDG has been described by Lawrence *et al.* [ref 22].

## **5. Discussion**

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 and the specificity of the data, it was practical to develop a metadata schema for ESG. The ESG prototype ontology has provided a framework for developing the ESG schema, highlighting the concepts of provenance and scientific investigation, expressing relationships such as PartOf and GeneratedBy, and separating domain specific concepts from more general ones. The

iterative work of detailed concept definitions and the rigor needed for specifying relationships between entities required in ontology authoring have served well to improve the schema.

ESG metadata is only partially based on semantics, and one challenge for ontology efforts is to devise mechanisms for inter-operability to other schemas. One possible solution would be to choose existing, suitable ontologies and provide mappings between the ESG schema and these ontologies. The NASA Jet Propulsion Laboratory earth sciences ontologies are under consideration for this purpose. Mappings will require that the ESG schema is represented in OWL, the W3C candidate standard language for ontologies. It is expected that challenges will arise in mappings related to time (simulation coverage, calendars) and spatial representation as geo spatial grids have numerous dimensions.

Semantics in NDG are relevant at three levels: metadata (for search and discovery), data (for virtualization), and services (for orchestration). The NDG metadata schema itself attempts to incorporate limited semantics. Relationships between the core entities (data activity, production tool, observation station) enable searching to cross these dimensions in a meaningful manner. Interoperability initially will be supported through export of metadata in standards-compliant formats (DIF, FGDC, ISO 19115 etc). A longer term plan to explore ontology-based interoperability will develop semantic mappings to other metadata models (ESG for instance).

In ESG, searches are more focused since the items satisfying search criteria are relatively few (compared to a Web search). An ESG search is more akin to a directory search than to a keyword search. However, the multiplicity of catalogs, their size, the volume of data indexed, the lack of information describing datasets and their content, and the hierarchy systems used by catalog implementations required solutions that benefited from Semantic Web technologies. The ESG prototype ontology attempts to clarify relationships between datasets, scientific investigation, and data pedigree for the ESG metadata schema. 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.

Not all metadata is used for discovery purposes. Only the simulation object is currently exposed to (free-text) searches. Keyword searches may return a schema element name or a value for an element without distinguishing them. For instance, a search on “ocean” may return a dataset where ocean is mentioned in a note, or is a type of simulation model. XML does not allow direct encoding of items in a set so that an ESG dataset metadata instance is linked to a parameter list, but not to items in the list. Searches on parameters that would return datasets containing these parameters are not currently implemented. A possible design for implementing parameter searches would include migrating relevant parts of the ESG schema to a new ontology and represent it in a language that permits searches on items in a set. Figures 9 show a representation of dataset instances in DAML and OWL. PCM.B06.10.dataset1 has parameters `bounds_latitude` and `cloud_medium` (Figure 9a), and PCM.B06.10.dataset2 has parameters `bounds_latitude` and `temperature` (Figure 9b). Figures 9a and 9b were created with OilEd 3.5 [ref 18].

## Figures 9a and 9b

Two key ESG contributions towards discovery services are the representation of collections of datasets, and the implementation of Logical and Physical file names. The benefit is the creation of a “virtual” dataset with its own unique metadata instance, such as for collections. Metadata is unique for a logical file, but applies to all the physical files pointed at by a logical file name. Metadata and Logical File Names are kept in the Metadata Catalog Service (Figure 8) and target locations of a physical file are kept in the Replica Location Service, a separate catalog (Figure 8), permitting the metadata to be easily browsed and searched independently of the physical files.

Metadata important for both projects appear similar in content but the paradigms under which they are organized have both gaps and overlaps. Federating schemas and sharing tools appear non-trivial. The ESG and NDG schemas describe entities directly related to

the Earth Sciences domain but pedigree classes are part of the ESG schema and belong to several schemas (NDG and CCLRC) in the UK system.

ESG metadata services are compliant to the Open Grid Service Infrastructure where a Grid service instance is a (potentially transient) service that conforms to a set of conventions for such purposes as lifetime management, discovery of characteristics, notifications and so forth<sup>[ref 19]</sup>. Services in ESG cannot be composed as they would be according to the Semantic Web vision. Several pre-defined workflows are possible in the ESG architecture but choices based on user preferences are not fully automated and workflows are currently hard-wired. Peer-to-peer interaction cannot be negotiated based on data content and rules as in some agent-based systems.

NDG is committed to a standards-based approach as far as possible. The ISO Technical Committee 211 is developing a range of standards for geographic information. Under this program, semantic data models are developed for a range of data types [ISO 19103, ISO 19109], and catalogued for re-use in endorsed registries (“feature-type catalogues”) [ISO 19110]. With respect to this program, the NDG data model may be considered an abstract feature type. Specialization will be undertaken as a community exercise. In addition, the emergence of standards-based registries will provide a core resource on which to develop ontology-based semantic mappings.

NDG service-level semantics are built from the semantic data model. Subsetting, for instance, will be based on conventional predicates applied to multidimensional arrays (e.g. start/stride/count subsampling), or filtering on spatial/temporal ranges. Specialization of the abstract model into domain-specific data types, in accordance with ISO standards, will enable sophisticated data services to be developed. For instance individual oceanographic profile data will be able to be aggregated and rendered as a vertical section through the ocean. In addition, data quality information [ISO 19113, ISO 19114] will be tailored as appropriate for different instrument types (expanding significantly the current “missingValue” flag used for model data).

## 6. Conclusion

This paper has described scientific data discovery for earth sciences data in the Earth System Grid and NERC Data Grid. Requirements, use cases, and challenges due to size, distribution of data, and poor annotations for earth science data were discussed. Solutions implemented by ESG included the ESG schema and discovery architecture. Metadata and search mechanisms for collections of data were implemented that constitute an important contribution to earth sciences data management. NDG solutions also included data and metadata hosts linked by services, and a metadata schema that separates concepts linked to provenance from those linked to data productions. The use of Semantic Web technologies such as ontologies has facilitated the development of the ESG schema, and opened the possibility of data interoperability on a federated basis, starting with the British collaboration. The targeted community of users was also discussed in the context of the Semantic Web. One primary concern was to enable these users to rapidly access, search and retrieve binary datasets from very large archives.

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 (although limited) language such as RDF is beginning to emerge in Grid communities. Methods for partial mappings and ontology reconciliation using pieces of common, small ontologies already exist and could be adapted for Grid purposes. The Earth System Grid provides the Semantic Web with testing grounds illustrating the complexity and magnitude of some scientific data problems. Interdisciplinary collaborations and the number of participants in scientific projects will only increase. The Semantic Web's focus on mechanisms for sharing information based on content, and tools for handling these complex tasks may bring a measure of relief to current obstacles in scientific grids. New developments in web service standards (the addition of RDF tokens to a WSDL service description, for instance) will enable orchestration of web and Grid services in a semantically intelligent

manner. NDG and ESG continue to follow Semantic Grid developments [ref 20] and will investigate potential as resources allow.

## References

- [1] Middleton, D. et.al. The Earth System Grid II: Turning Climate Datasets into Community Resources. AMS 2002. <http://www.earthsystemgrid.org/>
- [2] FGDC Metadata Workbook. [http://www.fgdc.gov/metadata/meta\\_workbook.html](http://www.fgdc.gov/metadata/meta_workbook.html).
- [3] W3C. [www.w3c.org](http://www.w3c.org).
- [4] Foster, I., Kesselman, C., eds. *The Grid: Blueprint for a new Computing Infrastructure, 2<sup>nd</sup> Edition*, San Francisco, Calif.: Morgan Kaufmann, Inc. 2004.
- [5] Foster, I., Kesselman, C., Tuecke, S. 2001. 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
- [6] CCLRC. <http://www.clrc.ac.uk/>
- [7] Buneman, P., Khanna, S. and Tan, W.-C., Why and Where: A Characterization of Data Provenance. *International Conference on Database Theory*, 2001.
- [8] The Global Change Master Directory. <http://gcmd.gsfc.nasa.gov/Aboutus/sitemap.html>.
- [9] The Dublin Core Metadata Element Set V1.1 (DCMES). <http://dublincore.org/usage/terms/dc/current-elements/>.
- [10] "Globus: A Metacomputing Infrastructure Toolkit," I. Foster, C. Kesselman, *International Journal of Supercomputer Applications*, 11(2):115-128, 1997; and <http://www.globus.org/>.
- [11] The Open Grid Services Architecture Data Access and Integration. <http://www.ogsadai.org/>.
- [12] 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.
- [13] Protégé-2000. <http://protégé.stanford.edu/>.
- [14] Peter N. Schweitzer (U.S. Geological Survey, Reston, VA 20192), Doug Nebert (USGS), Eric Miller, Quinn Hart, Jim Frew , and Archie Warnock. FGDC Metadata DTD Version 3.0.2, revised 2002-02-05. Available from

<http://geology.usgs.gov/tools/metadata/tools/doc/mp.html>.

[15] Foster, I., Kesselman, C., Nick, J. and Tuecke, S. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration Global Grid Forum, 2002.

[16] "High-Performance Remote Access to Climate Simulation Data: A Challenge Problem for Data Grid Technologies," W. Allcock, I. Foster, V. Nefedova, A. Chervenak, E. Deelman, C. Kesselman, J. Lee, A. Sim, A. Shoshani, B. Drach, D. Williams, SC'2001, ACM Press, 2001.

[17] Open Archives Initiative. <http://www.openarchives.org/>

[18] OilEd Editor, <http://oiled.man.ac.uk/>.

[19] Open Grid Service Infrastructure, GWD-R (draft-ggf-ogsi-gridservice-29), April 5, 2003. Page5.  
[http://www.gridforum.org/ogsi-wg/drafts/draft-ggf-ogsi-gridservice-29\\_2003-04-05.pdf](http://www.gridforum.org/ogsi-wg/drafts/draft-ggf-ogsi-gridservice-29_2003-04-05.pdf)

[20] O'Neill, K. et. al., The metadata model of the NERC DataGrid. In *Proceedings of the UK e-Science All Hands Meeting*. Nottingham, UK, September, 2003.

[21] Woolf, A. et. al., Data virtualisation in the NERC DataGrid. In *Proceedings of the UK e-Science All Hands Meeting*. Nottingham, UK, September, 2003.

[22] Lawrence, B.N. et. al., The NERC DataGrid prototype. In *Proceedings of the UK e-Science All Hands Meeting*. Nottingham, UK, September, 2003.

[23] Ramachandran, R. et. al., Earth Science Markup Language. *17th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography and Hydrology*. January, 2001.

[24] Ramachandran, R. et. al., Semantics and the Earth Science Markup Language. *Earth Science Technology Conference*, College Park, MD, Jun. 24 - 26, 2003.

[25] ISO 19101:2002, *Geographic information – Reference model*.

[26] The NetCDF Markup Language (NcML), <http://www.unidata.ucar.edu/packages/netcdf/ncml/>

[27] Data Format Description Language. <http://forge.gridforum.org/projects/dfdl-wg/>

## Figures

Figure 1: ESG overview.

Figure 2.:Conceptual dimensions required by NDG.

Figure 3: ESG prototype ontology classes.

Figure 4: Ontology relationships

Figure 5: Provenance information for dataset JDL\_00061.

Figure 6: ESG class diagram. (Courtesy of Bob Drach).

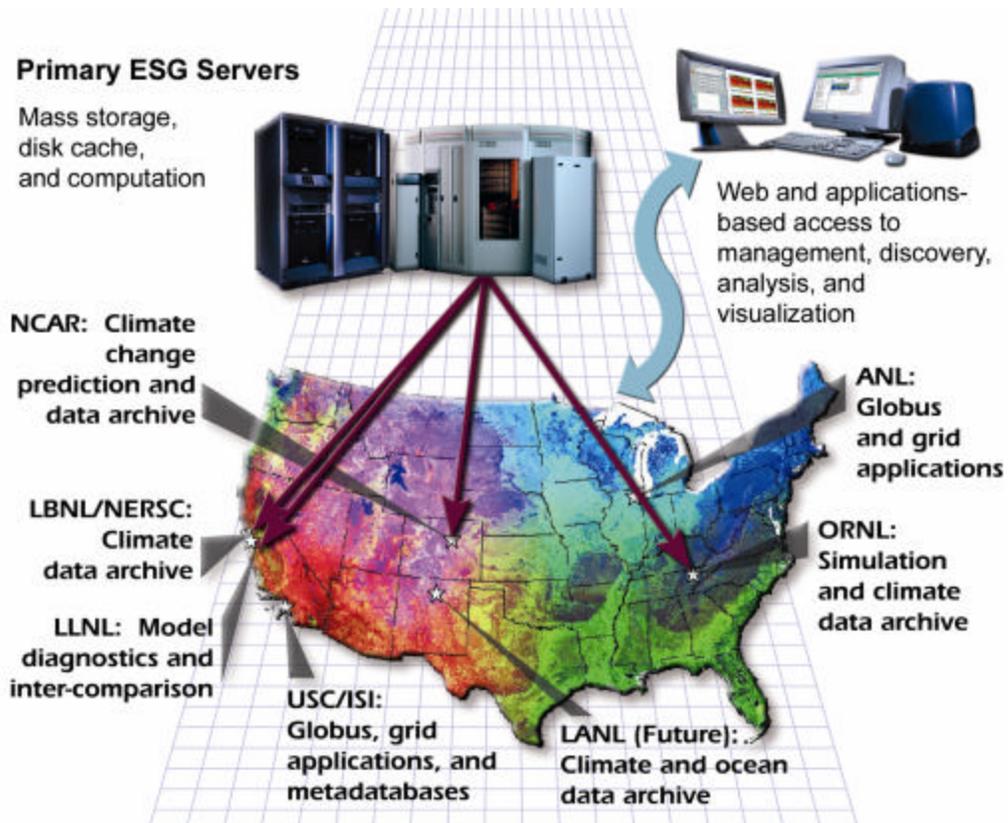
Figure 7: File name topology

Figure 8: ESG Discovery Architecture.

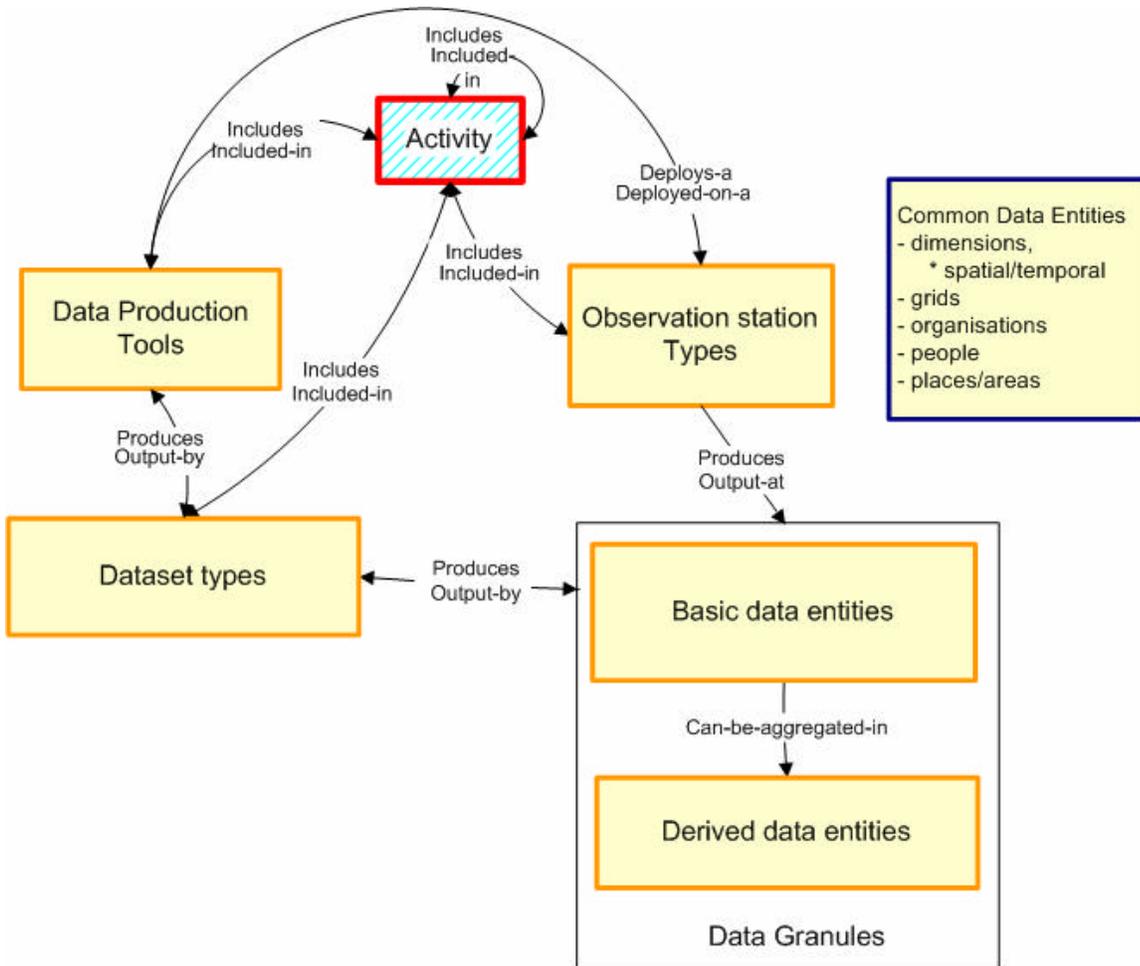
Figure 9a: Dataset instance representation with parameters bounds\_latitude and temperature in daml.

Figure 9b: Dataset instance representation with parameters bounds\_latitude and temperature in owl.

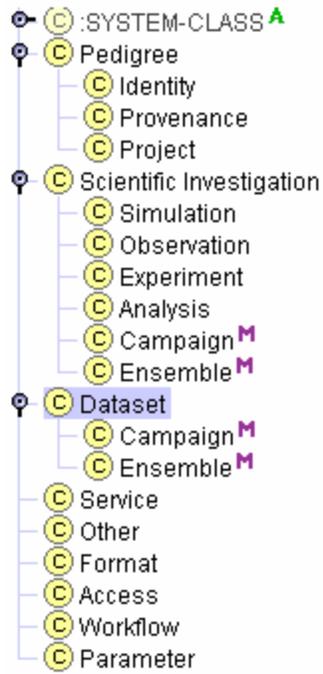
**Figure 1: ESG Overview**



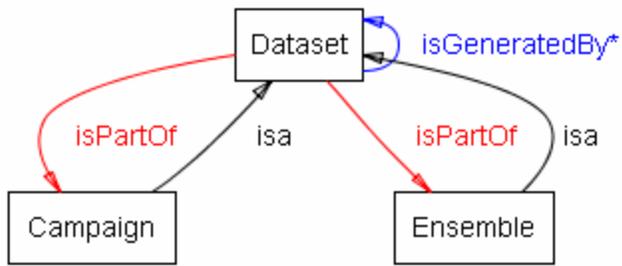
**Figure 2: Conceptual dimensions required in NDG.**



**Figure 3: ESG prototype ontology classes.**



**Figure 4: Ontology relationships.**



**Figure 5: Provenance Information**

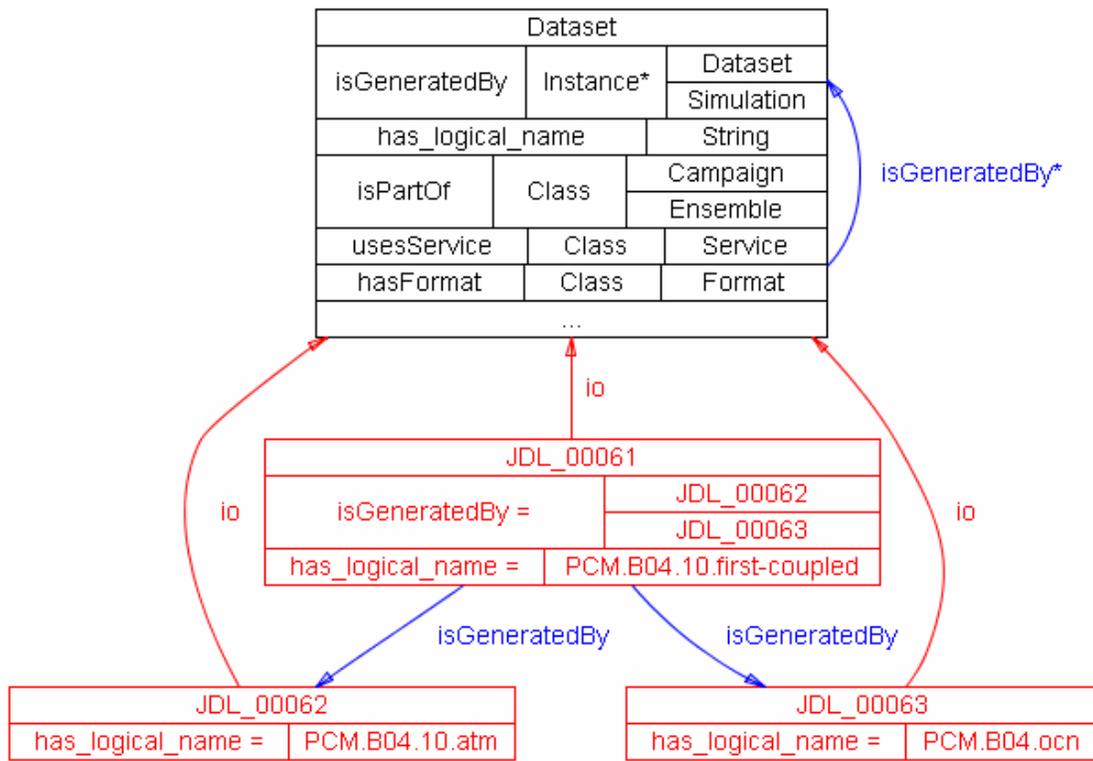
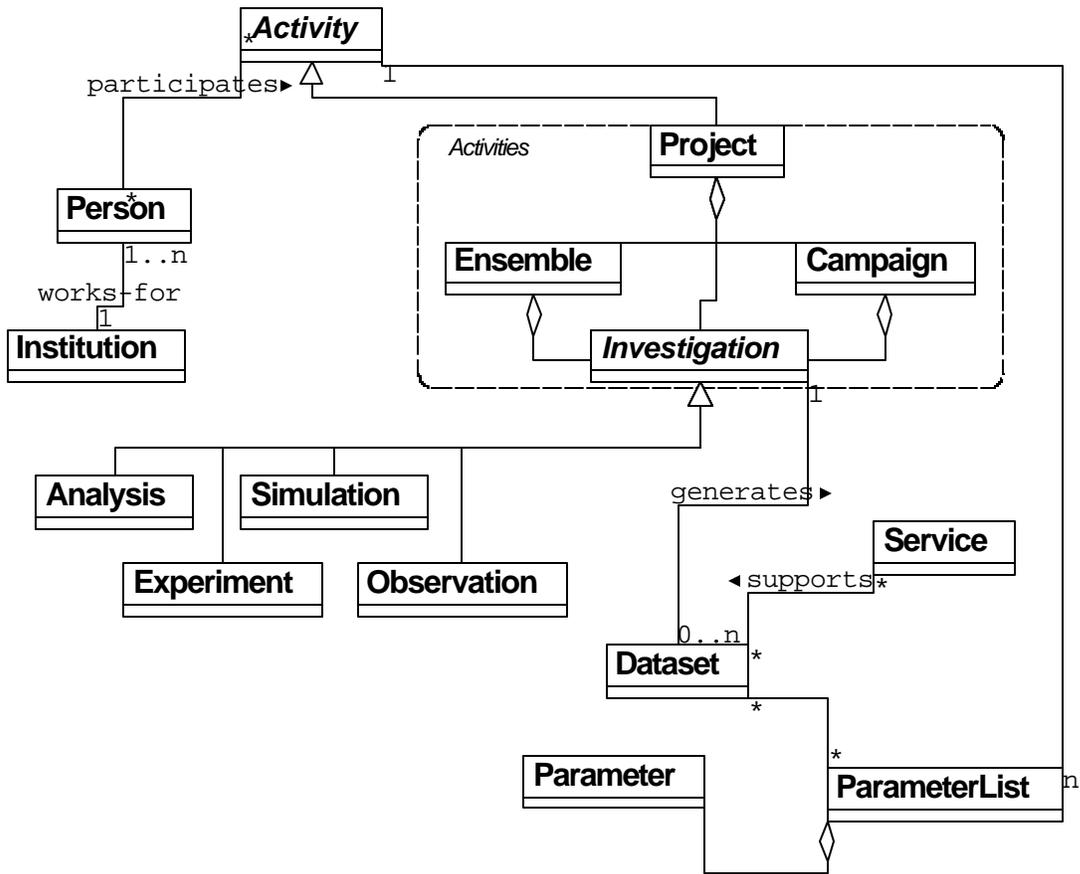


Figure 6: ESG Class diagram



**Figure 7: File name topology**

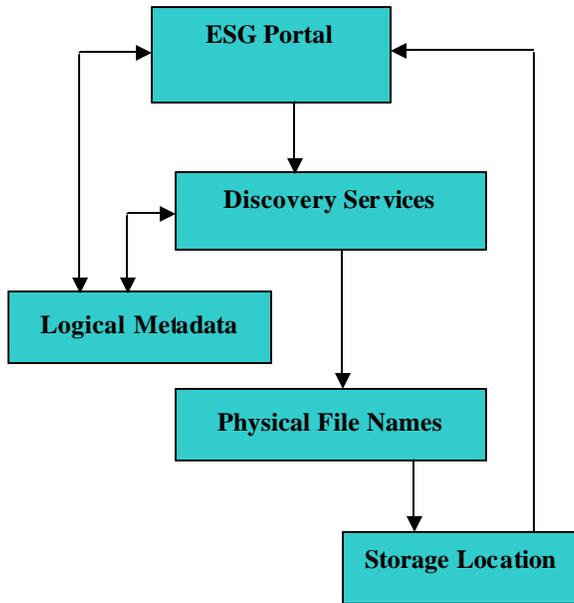
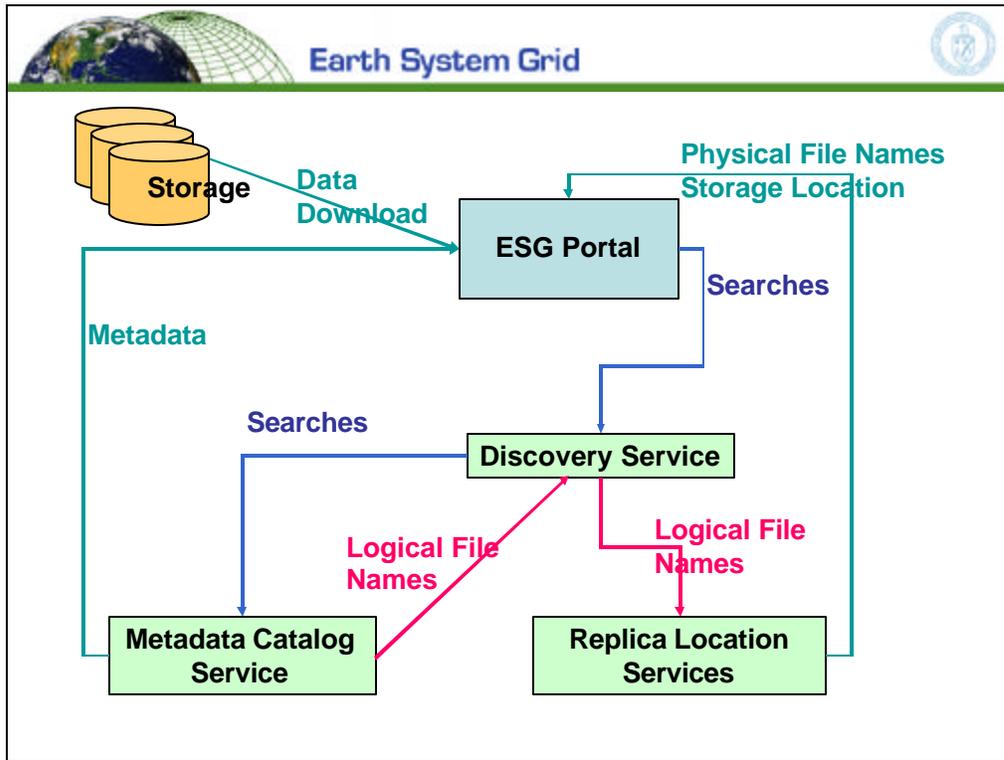


Figure 8: ESG Discovery Architecture



## Figures 9

**Figure 9a:** Dataset instance representation with parameters `bounds_latitude` and `cloud_medium`.

```
<rdf:Description rdf:about=" ../ESG/jdl_example.daml#PCM.B06.10.dataset1">
  <rdf:type>
    <daml:Class rdf:about=" ../ESG/jdl_example.daml#dataset"/>
  </rdf:type>
  <ns0:hasParameter df:resource=" ../jdl_example.daml#bounds_latitude"/>
  <ns0:hasParameter rdf:resource=" ../jdl_example.daml#cloud_medium"/>
</rdf:Description>
```

**Figure 9b:** Dataset instance representation with parameters `bounds_latitude` and `temperature` in owl.

```
<owl:Ontology rdf:about="">
  <dc:title>JDL_example</dc:title>
  <dc:date>December 09, 2003</dc:date>
  <dc:creator>Line Pouchard</dc:creator>
  <dc:description />
  <dc:subject>parameters in ESG schema</dc:subject>
  <owl:versionInfo />
</owl:Ontology>

<owl:Class rdf:about="file:/C:/Program%20Files/OilEd/ontologies/ESG/jdl_example.daml#dataset">
  <rdfs:label>dataset</rdfs:label>...</owl:Class>

<owl:ObjectProperty rdf:about="#jdl_example.damlhasParameter">
  <rdfs:label>jdl_example.damlhasParameter</rdfs:label>
  <rdfs:domain>
    <owl:Class>
  </rdfs:domain>
  <rdfs:range>
    <owl:Class>
    <owl:oneOf>
      <rdf:List>
        <rdf:first>
          <owl:Thing
rdf:about="file:/C:/Program%20Files/OilEd/ontologies/ESG/jdl_example.daml#cloud_medium" />
          </rdf:first>
        </rdf:List>
      </owl:oneOf>
    </owl:Class>
  </rdfs:range>
</owl:ObjectProperty>

...

<rdf:Description
rdf:about="file:/C:/Program%20Files/OilEd/ontologies/ESG/jdl_example.daml#PCM.B06.10.dataset2">
<ns0:hasParameter
rdf:resource="file:/C:/Program%20Files/OilEd/ontologies/ESG/jdl_example.daml#bounds_latitude" />
<ns0:hasParameter
rdf:resource="file:/C:/Program%20Files/OilEd/ontologies/ESG/jdl_example.daml#temperature" />
</rdf:Description>
```