10.13: Data Management Support for Adaptive Analysis and Prediction of the Atmosphere in LEAD

Beth Plale, Indiana University
Rahul Ramachandran and Steve Tanner, University of Alabama Huntsville

1. Introduction[1]

In 2003 the US National Science Foundation (NSF) funded a large information technology research (ITR) grant known as Linked Environments for Atmospheric Discovery (LEAD). A multidisciplinary effort involving nine institutions and more than 100 scientists, students, and technical staff in meteorology, computer science, social science and education, LEAD addresses the fundamental research challenges needed to create an integrated, scalable framework for adaptively analyzing and predicting the atmosphere. The high level goals of the project have been described in [4,5,10]. The specific research emphases are in dynamic and adaptive workflow using GBPEL[11], the myLEAD personal workspace [9], scientific portals[4], data mining[13], performance monitoring[1,2], and multiple resolution ensemble forecasts and dynamic adaptation techniques in meteorology forecasting[3].

With the dominant role played by data in all aspects of mesoscale meteorology, it is reasonable to expect that a large number of the requirements for the cyberinfrastructure in development in LEAD will be oriented towards data management support. For instance, forecast models are computationally intense, often requiring exclusive use of hundreds of processors for up to twelve hours. In order to kick off a forecast model on-demand in response to a severe weather event, one needs either exclusive access to a high-end cluster or supercomputer, a costly proposition, or access to a shared high-end compute resource with an agreement that that the resource will be available the moment it is needed. But the forecast models are data driven, in that they take initial conditions from observational and model generated data, so the cyberinfrastructure must, in addition to supporting real-time scheduling of the forecast model on a supercomputer resource, support automated search, selection, and movement of the appropriate data products needed by the forecast.

As another example, as the availability of large-scale shared compute resources such as the Teragrid grows, meteorologists respond with larger, more complex models carried out on a smaller grid spacing or nesting within a larger model, one or more models with finer grid resolution. This confluence of increased model complexity and automated model execution, results in exponential growth in the sheer volume of data products that must be managed. Under these circumstances, tracking, moving, and searching for data products exceeds any single scientist's ability to manage with the paltry tools provided by his/her desktop machine and local file system, that is, long file names, directories, and 'grep'. Finally, as anticipated in the latter years of the project, the forecast models will be able to ingest data streams at any point during execution directly from real-time sources, such as the CASA NETRAD radars, so the cyberinfrastructure must be able to route selected stream data from its source to the model wherever the model currently happens to reside.

The meteorology community has benefited from a relatively long history of access to a large number of observational and model generated data products, for instance, GOES satellite data, upper air balloon (Rawinsondes), ship and buoy (METAR) data, Nexrad Level II and III Doppler radar data to name a few. The relatively long time over which these products have existed and the general agreement by the community as to their value have resulted in the early establishment of community-supported data dissemination, access, and visualization tools. These tools, most notably Internet Data Dissemination (IDD), THREDDS, and IDV developed by Unidata serve a broad community of users. IDD efficiently routes observational and model data to any client machine that has the open source client installed, and THREDDS is an XML-based web server providing download access to data products.

With this strong existing foundation, what then is needed in the way of data management tools and

[1] LEAD is funded by the National Science Foundation under the following Cooperative Agreements: ATM-0331594, ATM-0331591, ATM-0331574, ATM-0331480, ATM-0331579, ATM03-31586, ATM-0331587, and ATM-0331578.

functionality to support the paradigm shift to integrated, scalable framework for adaptively analyzing and predicting the atmosphere that LEAD envisions? The data subsystem challenges being explored by LEAD as needed to satisfy adaptive analysis and prediction fall into three categories:

Automated data discovery – what we as computer scientists refer to colloquially as running a "weather forecast" is actually a complex sequence of steps including gathering data products, setting configuration parameters, assimilating the products into a single 3D volume, executing the model, and generating resulting products that are then analyzed by a statistical tool or visualized and analyzed by a human. This sequence, which we depict as *<data → model → analysis → results>* is called a *workflow*. In order for a weather forecast workflow to be kicked off and execute automatically in response to early severe storm conditions, it is necessary to replace the manual tasks of data management with automated ones. This means that searching for input data products needed by a workflow, and capturing and storing the output data products for a user must be automated.

Highly scalable data archiving system – by introducing automated workflows as the means by which forecasting is done, this opens the opportunity to scale the forecast model to levels well beyond what is done today. Nested forecast models, where smaller grid spacing is nested within larger grid spacing, and ensemble models where 100-500 models are simultaneously executed, now fall into the realm of reality. The data management challenges to support the scale of forecasting envisioned requires considerable attention to movement and storage of terabytes of data. No longer is it possible for a single user to organize on his/her own workstation all the data products generated during the runs. Storage facilities located on the computational grid need to be available to a user, providing the same guarantees of privacy and protection as his/her own file system.

Easy search and access to data – not every step of the forecast can be automated. The user must still indicate the starting conditions and specify the parameters of the run. But today this task is exceeding difficult because it requires significant expertise to know what data products contain what kinds of data, where the products are located, and how they are to be used. In LEAD we are easing the task by providing a search GUI, ontology, and search services to ease the task of locating data products. The solution we are exploring is general. That is, additional data collections, with formats not yet known to us, should be able to be added to the system and discovered as easily as the known data collections are today. The importance of this feature will become obvious when researchers move on to coupling an atmospheric model with a hydrological model, or some similar cross-discipline coupling.

In this paper we discuss three recent developments of the data subsystem that our groups are prototyping as solutions to one or more of the goals identified above. These are a metadata representation based on the FGDC standard, the OIS ontology, and the myLEAD personal workspace. These three developments in the LEAD data subsystem are key early outcomes of the ongoing fundamental research in creating an integrated, scalable data management framework for adaptively analyzing and predicting the atmosphere.

2. Metadata in LEAD Data Subsystem

Metadata is generally defined as "data about data". The dictionary defines meta as "beyond, transcending, more complete". In the LEAD context then, metadata is information about a resource where the resource can be information, datasets, workflow or compute resources. LEAD is a complicated system and without such metadata, its resources can become difficult to harness and lost in the noise of too much information. Thus, metadata is the key to ensuring that resources in a project survive and continue to be accessible and utilized in the future.

Formal metadata is metadata that follows a standard specification that provides a common set of terminology, definitions and information about the values to be provided. Such metadata are a formally structured documentation of resources, describing the who, what, where, when, why and how of every aspect of the resource. It is useful in organizing and maintaining an organization's internal investment in a resource. It provides information to data catalogs, clearing houses, search engines and can form the information currency that is exchanged between different components within and outside the system. A formal metadata approach is essential for scientific data and projects.

A different approach is used by search engines such as Google or Alta Vista where every word in a document is indexed, thus harvesting metadata without using a formal specification. This approach is applicable to documents but not to science data. Furthermore, these methods used for harvesting metadata have limitations in the accuracy of their results. Thus researchers are looking at approaches such as the Dublin Core Initiative [http://dublincore.org/] to formalize presenting additional information relating to the documents.

2.1 Role of Metadata in LEAD

Metadata in LEAD plays a crucial role in three areas. These are:

1. Facilitate Discovery and Access of LEAD Resources
Metadata will help describe content information to allow resource discovery through either a query service or via the semantic search engine. It will also provide location information for resource access.
2. Facilitate Use of LEAD Resources
Metadata will provide syntactic and semantic information for resource interoperability and integration.
3. Facilitate Preservation of LEAD Resources
By storing information such as quality, provenance, etc., metadata will ensure digital identification and preservation.

2.2 Metadata Design Principles

The design principle of 3W's was used to design the LEAD metadata. The 3W's represents the three key questions Who, What and Why that the metadata design must address.

Who is the metadata for?
Before designing the metadata, it is important to understand the different users of the metadata and their specific needs. For the LEAD project, these users can be students from high school to graduate level, teachers at these levels and finally, atmospheric science researchers.

What metadata standard should one use?
There are several metadata specifications that one can use to describe resources such as geospatial datasets. One has to carefully select the specification that covers the needs of the target users of the metadata and permits interoperability with other systems using different specification. The LEAD team has selected Federal Geographic Data Committee (FGDC) [http://www.fgdc.gov/metadata/metadata.html] standard as the basis for its metadata, and has tailored it to fit their specific needs.

Why use specific metadata elements?
Metadata standards such as FGDC are extremely broad and try to cover every aspect of the resources described. To effectively use these standards, one has to create "profiles" of these standards to suit the project requirements. Creation of these profiles requires deleting, modifying or adding new metadata elements based on its importance to the project needs.

2.3 LEAD Metadata Schema Overview

A FGDC profile for the LEAD project was created using these principles. A high level overview of the LEAD metadata schema is presented in Fig 1. The current schema was designed for scientific datasets and grid workflows as its target resources. Each resource is assigned a unique *resource id* for identification purposes. The metadata describing the scientific datasets contains two mandatory components, the *idinfo* and the *metainfo*.

The *idinfo* element covers all the basic information required to identify the datasets. It covers the following mandatory metadata elements:

- *Citation*: captures information such the name of an organization or an individual that developed the dataset, the date when the data set was released and the name by which the data set is known.

- *Description*: captures a brief narrative summary of the dataset and the summary of intentions with which the dataset was created.

- *Status*: captures the state of the dataset along with the frequency with which changes and additions are made to the dataset after the initial dataset is completed.

- *Access Constraints*: captures any restrictions and legal prerequisites for accessing the data set. These include any access constraints applied to assure the protection of privacy or intellectual property, and any special restrictions or limitations on obtaining the data set.

- *Use Constraints*: captures any restrictions and legal prerequisites for using the data set after access is granted.

- *Keywords*: captures the words or the phrases summarizing an aspect of the data set. These include subjects covered by the data set.

The *metainfo* element provides details about the metadata reference information. It captures information on the currentness of the metadata information and the responsible party.

The optional elements in the LEAD metadata schema include:

The *distinfo* element provides distribution information such as information about the distributor and options for obtaining the data set.

The *dataqual* element provides a general assessment of the quality of the data. It also captures the lineage information.

The *geospatial* element has been specifically added to the LEAD metadata schema by restructuring the original FGDC specification. It covers both the spatial and temporal coverage of the datasets. This restructuring was essential as the LEAD project will also generate datasets with no spatial or temporal component.

The *enclosedresouces* element is another component specific to LEAD that has been added to the FGDC schema. This element allows us to capture the notion of data collections or aggregations that is missing in the original FGDC specification.

The basic information covered in the *idinfo* element for datasets is also used to describe the workflows.

Note that the restructuring of the FGDC specification to meet the LEAD requirements was done in keeping with the overall FGDC specification spirit. All the mandatory elements were kept in this profile and only the optional elements were discarded or modified.

3. LEAD Ontology

The use of a standard metadata specification provides the blueprint for the metadata elements and their definitions that can be used in a project. However, one other question still needs to be addressed: What actual values should be used to populate the metadata schema? For certain metadata elements such as time or spatial bound this is not a big issue as one could use ISO or IEEE standards to describe date or time. However, it is a vital design issue for abstract elements such a keywords. These elements are frequently used to search metadata catalogs to find the correct datasets. There are two approaches to address this question. The first approach enforces a control vocabulary. This means everyone, including the users in the project will use a known set of keywords. There are several control vocabularies in Earth science such as the Global Change Master Directory (GCMD) [http://gcmd.gsfc.nasa.gov/], the Climate and Forecast (CF) [http://www.cgd.ucar.edu/cms/eaton/cf-metadata/] Metadata Convention, etc., defined by different groups.

The drawback of using a control vocabulary is that everyone in the project must know all the keywords in the set. Such an approach is not very practical for LEAD because of the different levels of expertise of the end users. The end users of the LEAD system can vary from advanced researchers to sixth grade high school students. The second approach of using ontologies can address this problem, and provide an elegant and extensible solution.

An ontology has been described by Aristotle as the science of being. From Machine Learning/Artificial Intelligence/Intelligent Systems perspective "an ontology is a formal, explicit specification of a shared conceptualization" [7]. Therefore, an ontology contains concepts and constraints of use that are explicitly defined. It is formal, making it machine understandable and it is shared, meaning that it captures consensual knowledge. An Ontology tends to have two distinct components. It contains the names for important concepts for a specified domain. In addition to definition of concepts, the ontology also specifies the background knowledge and the constraints of the domain. Thus, an ontology cannot only act as an extended control vocabulary but also provides the context and relationships for the values.

3.1 LEAD Ontology

The LEAD ontology is being built using the Semantic Web for Earth and Environmental Terminology (SWEET) ontology [12]. SWEET is based on the NASA Global Change Master Directory (GCMD), which includes approximately 1000 controlled Earth science keywords, represented in a taxonomy. SWEET has been designed as a higher level ontology allowing domains within Earth Science to create specialized ontologies leveraging the SWEET concepts. The LEAD ontology will focus on concepts relevant to meteorology. As part of LEAD, the American Meteorological Society glossary is being mapped into the SWEET ontology by using the concepts listed in the glossary and defining relationships between them. In addition to the AMS glossary, terms used in the NetCDF Climate and Forecast (CF) convention are also being mapped into SWEET. Therefore, the LEAD ontology will be a specialized ontology for Meteorology. It will extend the concepts defined in the SWEET ontology and act as a superset of both the GCMD and the CF control vocabulary.

3.2 Ontology Inference Service

An ontology of course does not exist in a vacuum. The LEAD team is developing tools which make the ontology available to both users and other services within LEAD. Primary among these is the Ontology Inference Service (OIS). OIS is a SOAP-based web service interface to an inference engine. It is built on the Apache Axis SOAP engine. The inference engine used at the backend is Pellet[6], an OWL DL reasoner based on the tableaux algorithms. The reasoner is pre-loaded with the LEAD ontology and provides T-Box and A-Box querying capabilities on the ontology. T-Box queries cover

specializations, generalizations and equivalences of a concept. A-Box queries search for all satisfying instances of a concept and for property fillers for an instance. Every search request to the OIS is translated to one or more such queries for the reasoner. The OIS interacts with the reasoner through the description logic reasoner interface (DIG). The DIG interface is a standard for providing access to description-logic reasoning through an HTTP-based interface. The query results are returned back to the OIS through this interface. OIS has been integrated with the LEAD query service to provide ontology-based semantic search capabilities. These search capabilities include both a Yellow Page Search and Synonym Matching. The Yellow Page Search will allow a user to search by providing a higher level concept such as Temperature and the OIS will find specializations of this concept and return those terms to the LEAD query service to locate the appropriate datasets. The OIS will also be able to find synonyms for the search terms to ensure accurate and complete search results.

4. Personal Workspace

Scientists have long had to deal with managing the derived data products from their experiments with minimal tool support. For example, forecast models such as ARPS accept assimilated observational and model data as input but the task of moving the data products to the computation is largely a manual process. Similarly, results have to be moved off the large compute resource back to the scientist's institution. In some systems this process can be automated, that is, handled by large, brittle scripts. Even if automated, though, some months or years in the future a scientist searching his or her directory tree for a specific file or set of results will likely have difficulty finding a particular file. Much of the data written to long-term store is never accessed again.

The personal workspace, shown as *myWorkspace* in Figure 2, is a cornerstone in managing a user's scientific data. It provides a user of the LEAD grid with a persistent private space for his/her model results and a host of other information related to his/her investigations.. The user interacts with his/her workspace through the LEAD portal gateway. Access to the portal and tools is controlled by a X.509 certificate security scheme. Once logged in, the user can search or browse their workspace. The myExperiment space is a set of tools and interfaces for building and running workflows. The myWorkspace works in conjunction with the Experiment Builder tools to locate and stage data products in anticipation of an execution.

The infrastructure supporting the myWorkspace concept goes well beyond client-side GUI's and tools, however. As shown in Figure XX, the data management infrastructure supporting the myWorkspace include the OIS ontology service described earlier, a community resource catalog, and the myWorkspace catalog. The community resource catalog, not the subject of this paper, is simply a registry of community resources, including services and data products. An intuitive view of the catalog in the way it handles data products is as a centralized Apache Lucene index over a set of THREDDS catalogs. It provides additional benefit over direct integration of THREDDS catalogs in that it "talks" the LEAD metadata schema that is the *lingua franca* in which services talk to one another about data products.

The myWorkspace middleware separates the storage of the metadata from the data products. The limitations of a long filename, unique UNIX path name, and 'grep' are widely known. To enable richer search, data products must be described by these important attributes, but application-domain attributes as well. For instance, the metadata for a Doppler radar observational scan could include an instrument's description, the starting time of the scan, instrument type and spatial location, or unique four-letter mnemonic.

4.1 Service Architecture

The myWorkspace middleware (also called "myLEAD") is anticipated to support hundreds of active personal workspaces simultaneously. The architecture can be viewed as a set of distributed services that cooperate with one another to give the user the impression that they are working with a single centralized service. An instance of the MyLead metadata catalog resides at each site in a grid testbed. Specifically, each of the five sites in the LEAD testbed will run a persistent server-side service that manages the personal metadata catalogs for users local to that site. A storage repository will reside at two sites, and will be used to store the files themselves. The user interacts with the myLEAD service through the LEAD portal, which is web accessible from anywhere on the Internet.

The server-side catalog is a persistent web service built on top of a relational database. It extends the Globus Toolkit Metadata Catalog Service (MCS)[15] and the Open Grid Services Architecture Data Access and Integration (OGSA-DAI) grid interface layer[8]. MyLead extends and improves on MCS through support for spatial and temporal attributes, significantly more improved query access, contextual queries, and support for LEAD

metadata.

The storage repository, shown in the lower left of Figure 3, could be as simple as a local file system, but storage repository solutions—the Globus Toolkit Distributed Replica Service (DRS), Storage Resource Broker (SRB)[14], and Storage Resource Manager (SRM)[16], for example—provide additional abstractions beyond a file system API, such as a notion of a container, location-transparent data storage, and global naming. Unidata is currently developing a storage repository solution that has many of the same features as DRS, but integrates support for OPeNDAP and the relatively new Unidata Common Data Model. Although the metadata catalog of myLEAD could interoperate with any of these repository tools, SRM and SRB tightly couple their own metadata catalog to their storage system, which introduces a redundancy that could have costly performance implications.

4.2 myLEAD Agent: Adding Value to myWorkspace

Though a user's workspace, when viewed from the inside is strewn across a dozen or more tables in a database, through the addition of an agent service layer, we give the user and the programs executing on behalf of the user, a hierarchical view of their space. It is well known that the hierarchical organization of information is intuitive for humans. In addition to displaying the information hierarchically in the portal for a user to browse, the myLEAD agent works on behalf of a user during an experimental investigation to track the different modes, or states, of workflow execution (e.g., 'model input state', 'model execution state'). It uses this knowledge to actively organize the metadata into named buckets corresponding to that state. These named buckets can then be tied to user concepts through the OIS ontology service. Users can then issue queries not only on the atmospheric terms, but also on stages in the investigation. As an example depicted in Figure 3, on December 2004 Bob has three experiments in his workspace. The study named "Vortice Study '98-00" has 3 collections, one for each of input products, workflow products, and model outputs. The outputs of his workflow are 150 NetCDF files. As of February 2005, Bob has run 2 archive-worthy experiments. Note that the hierarchy under "Vortice Study '98-00" has been automatically extended to capture both historically vital versions of this run. On March 2005, Bob is happy with the results from the February 2005 run and publishes select products from the experiment to the broader community of researchers.

5.0 Conclusion

In this paper we focus on the services, functionality, and tools needed to support the major paradigm shift LEAD is brings to mesoscale meteorology by means of an integrated, scalable framework for adaptively analyzing and predicting the atmosphere. The data subsystem challenges explored in this paper are in direct support of the larger goal. These include support for *automated data discovery* that is, replacing the manual tasks of data management with automated ones, *highly scalable data archiving system,* that is, movement, metadata description, and organization of terabytes of data, and user protected storage facilities located on the computational grid that provide the same guarantees of privacy and protection as does a user's own file system. Finally, we are providing *easy search and access to data* for easing the task by providing a search GUI, ontology, and search services to ease the task of locating data products.

An alpha version myLEAD was released May 2005; version 1.0 is slated for release by the end of 2005. The LEAD portal is accessible off the LEAD project page (http://lead.ou.edu).

6.0 Acknowledgements

The authors wish to thank fellow LEAD data-thrust group members for hours of stimulating discussions on data related topics, in particular, Jay Alameda, Tom Baltzer, Doug Lindholm, and Anne Wilson. The authors are deeply indebted to the many people involved in these projects, including Dr. Sangmi Lee Pallickara and students Yogesh Simmhan, Yiming Sun, Sunil Movra, Scott Jensen, and Ning Liu.

References

[1] Blatecky, A., K. Gamiel, L. Ramakrishnan, D. Reed, and M. Reed, "Building the Bioscience Gateway, Science Gateways: Common Community Interfaces to Grid Resources," *presented at Global Grid Forum*, Chicago IL, 2005.

[2] DeRose, Y. Z. Luiz, and Daniel A. Reed, "SvPablo: A Multi-Language Performance Analysis System," *10th International Conference on Computer Performance Evaluation - Modeling Techniques and Tools - Performance Tools '98*, Palma de Mallorca, Spain, 1998.

[3] Dietachmayer, G. and K. Droegemeier, "Application of continuous dynamic grid adaptation techniques to meteorological modeling, Part1: Basic formulation and accuracy," *Mon. Wea. Rev.*, vol. 120, pp. 1675-1706, 1992.

[4] Droegemeier, K., *et al.*, "Service-oriented environments in research and education for dynamically interacting with mesoscale weather," *IEEE Computing in Science and Engineering (CiSE)*, Vol. 7, No. 6, Nov/Dec 2005.

[5] Kelvin K. Droegemeier, V. Chandrasekar, Richard Clark, Dennis Gannon, Sara Graves, Everette Joseph, Mohan Ramamurthy, Robert Wilhelmson, Keith Brewster, Ben Domenico, Theresa Leyton, Vernon Morris, Donald Murray, Beth Plale, Rahul Ramachandran, Daniel Reed, John Rushing, Daniel Weber, Anne Wilson, Ming Xue, Sepideh Yalda, Linked environments for atmospheric discovery (LEAD): Architecture, Technology Roadmap and Deployment Strategy, *21st Conf. on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology*, January 2005

[6] Grau. C., B. Parsia, and E. Sirin, "Tableau Algorithms for E-Connections of Description Logics," University of Maryland Institute for Advanced Computer Studies (UMIACS) Technical Report, 2004.

[7] Gruber, T. R., "A Translation Approach to Portable Ontology Specifications," *Knowledge Acquisition*, vol. 5, pp. 199-220, 1993.

[8] OGSA-DAI Open Grid Services Architecture Data-Access and Integration http://ogsadai.org.uk

[9] Plale, B., D. Gannon, J. Alameda, B. Wilhelmson, S. Hampton, A. Rossi, and K. Droegemeier, "Active Management of Scientific Data," *IEEE Internet Computing special issue on Internet Access to Scientific Data*, vol. Vol.9, No.1, pp. pp. 27-34, 2005.

[10] Plale, B., D. Gannon, D. Reed, S. Graves, K. Droegemeier, B. Wilhelmson, and M. Ramamurthy, "Towards Dynamically Adaptive Weather Analysis and Forecasting in LEAD, *ICCS workshop on Dynamic Data Driven Applications*, Lecture Notes in Computer Science, Part II, LNCS 3515, Springer Verlag, 2005.

[11] Plale, B., D. Gannon, Y. Huang, G. Kandaswamy, S. L. Pallickara, and A. Slominski, "Cooperating Services for Managing Data Driven Computational Experimentation," *IEEE Computing in Science and Engineering (CiSE)*, Vol. 7, No. 5, Sep/Oct 2005.

[12] Raskin, R. G. and M. J. Pan, "Knowledge representation in the semantic web for Earth and environmental terminology (SWEET)," *Computers & Geosciences*, vol. 31, pp. 1119-1125, 2005.

[13] Rushing, J., S. J. Graves, E. Criswell, and A. Lin, "A Coverage Based Ensemble Algorithm (CBEA) for Streaming Data," *IEEE Intl. Conference on Tools with Artificial Intelligence*, Boca Raton, FL, 2004.

[14] A. Shoshani, A. Sim, and J. Gu, Storage Resource Managers: Middleware Components for Grid Storage, *IEEE Conference on Mass Storage Systems and Technologies* (MMS) 2002

[15] G.Singh et al. A Metadata Catalog Service for Data-Intensive Applications. *ACM/IEEE Supercomputing* 2003, IEEE CS Press, 2003, pp. 33—49

[16] SRB Storage Resource Broker, http://www.sdsc.edu/srb/

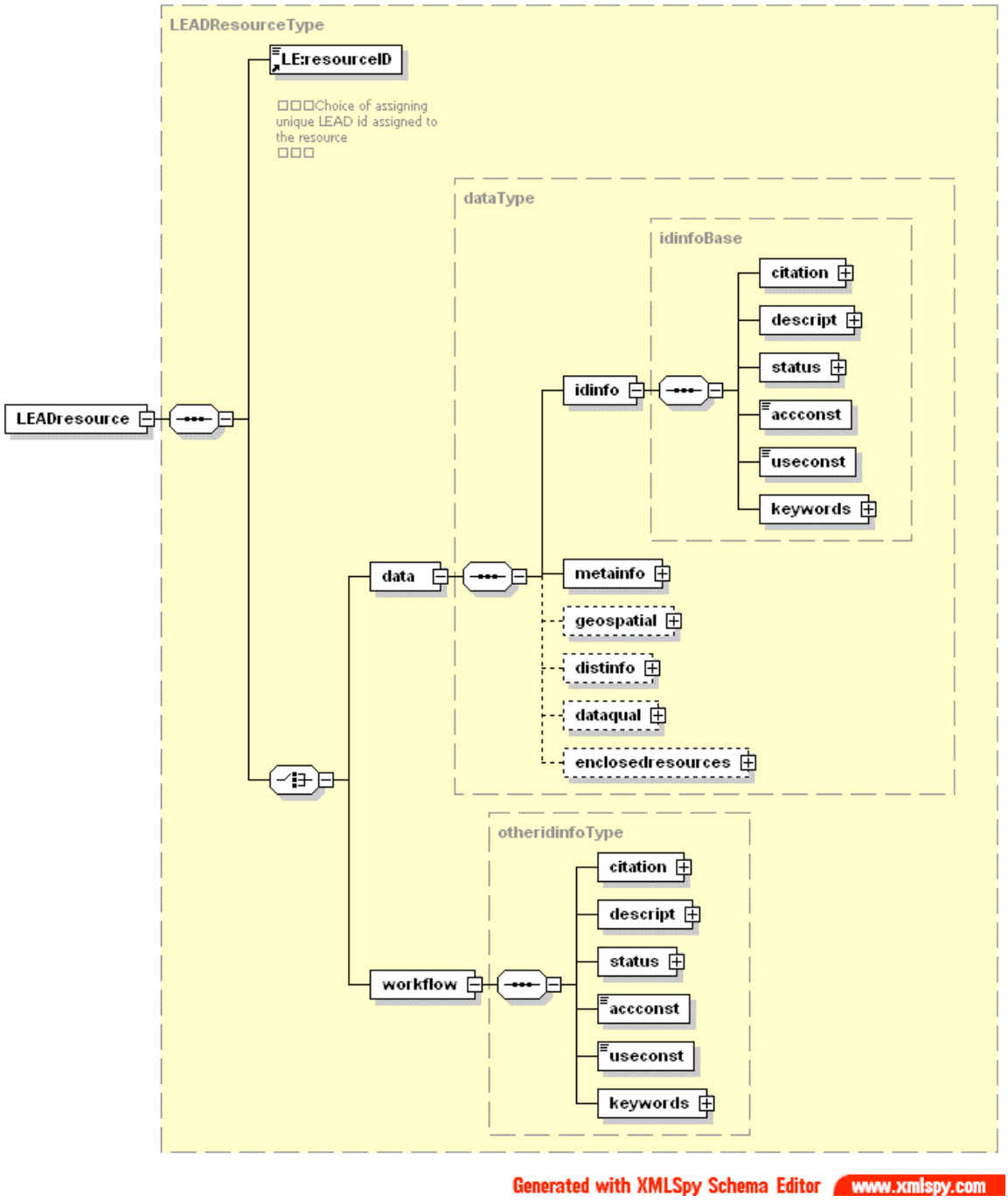

Figure 1: LEAD Metadata Schema

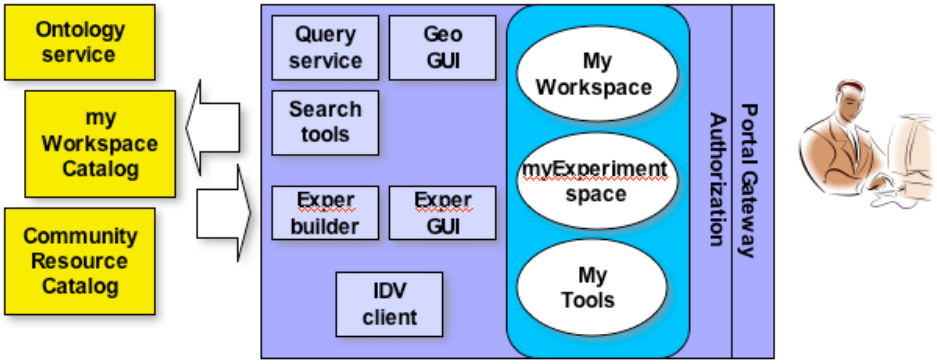

Figure 2. User access to LEAD data management services is through the LEAD portal. Client-side services and tools create support conceptual spaces in which the user works: myWorkspace, myExperiment, and myTools. The back end data services include metadata catalogs for personal and community resources and an ontology.

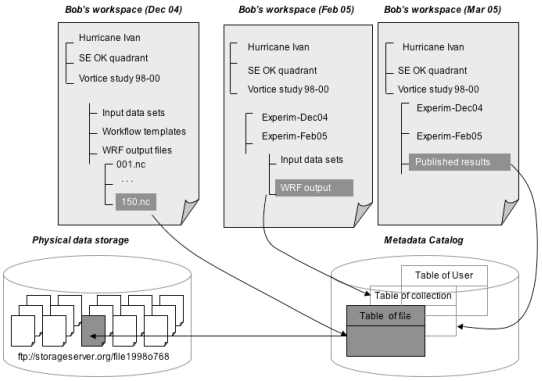

Figure 3. myWorkspace is supported by multiple services, of which two are shown in this figure. The metadata catalog stores metadata descriptions separate from the data products themselves in a database. The files are accessed separately using an ftp, http, or OPeNDAP access protocol. The user browses and searches his/her workspace through the LEAD portal.