5.8 A WEB-BASED SCIENTIFIC DATA AND INFORMATION SUPER SERVER WITH A FLEXIBLE XML METADATA SUPPORT

Ruixin Yang[*], X. Sean Wang, Yixiang Nie, Yujie Zhao, Menas Kafatos
Center for Earth Observing & Space Research (CEOSR)
George Mason University (GMU)

ABSTRACT

In the information age and with explosively increasing volumes of remote sensing, model and other Earth Science data available, scientists are now facing challenges to find and to access interesting data sets effectively and efficiently through the Internet. In this paper, we first discuss the DIstributed MEtadata Server (DIMES) prototype system. Designed to be flexible yet simple, DIMES uses XML to represent, store, retrieve and interoperate metadata in a distributed environment. In addition to regular metadata search, DIMES provides a web-based metadata navigation interface by using the "nearest neighbor search." We also discuss a system designed to integrate an existing data access and analysis server, namely the GrADS/DODS server, and the DIMES to form a Scientific Data Information Super Server (SDISS), which supports both metadata and data. The SDISS guarantees the consistency between the content of the data server and the content of the metadata server. A use case of the SDISS is also discussed.

1. INTRODUCTION

Applications and products of Earth observing and remote sensing technologies have been shown to be crucial to our global social, economic, and environmental well being. Earth observing from space produces very large amounts of data at ever-expanding rates. The Earth Observing System (EOS) satellite *Terra* alone is adding more than half a terabyte of Earth science data each day [Asrar & Greenstone 1999], and other Earth observing platforms and computer weather and climate models are producing or will produce even more massive data products. The volume and complexity of this information will only continue to grow, together with the stringency of the requirements/usage by different communities and in different environments. Using traditional data access and analysis methods, Earth science researchers will have increasing difficulties in handling, retrieving, analyzing, and presenting such massive amounts of data. Therefore, distributed data information systems with effective search capability are needed to help scientists to perform their research more efficiently.

The ultimate goal of the data information systems is to make the data available to scientists. In this information age, that means scientific users can find data, evaluate data and access and use all related data online regardless of data locations and formats. This requires high-level data interoperability including data delivery through the Internet. A data delivery mechanism as simple as FTP is very useful for data exchange although its limitation is obvious. High-level data delivery systems have been developed in different communities. A successful example is the DODS (Distributed Oceanographic Data System) originating in the oceanography community [DODS 2002]. Recently, an enhanced version of DODS by combining a data analysis capability of GrADS (the Grid Analysis and Display System) [Doty *et al.* 1997], GDS (GrADS/DODS Server) [Doty *et al.* 2001], even allows users to define operations performed on the server and to

[*] Corresponding author address: Ruixin Yang, MS 5C3, School of Computational Sciences, George Mason University, Fairfax, VA 22030; e-mail: ryang@gmu.edu.

obtain the resultant information (processed data) via the Internet. GDS, though powerful for data services, does not have enough searchable metadata for users to locate data quickly.

How to find useful data sets through the increasing available data in the electronically connected world is a difficult task. Commonly, metadata are used to search data sets and to find other useful information for obtaining and using the data. A major problem with the current metadata systems serving the Earth science community is that the search engines, usually large and supported by national centers, contain too much information for serving everyone so that there are too many hits for a specific search and some of the data links may be out of date. Under the Seasonal to Interannual Earth Science Information Partnership program [Kafatos *et al.* 1997], we have developed an XML-based DIstributed MEtadata Server (DIMES) [Yang *et al.* 2001] comprising a flexible metadata model, search software, and web-based interfaces to support various level metadata accesses. The major difference between our metadata model and others is that we kept small data providers in our mind during the development and therefore designed a flexible system, easily tailored to specific use and integrated with other systems.

To overcome the bulkiness problem of central metadata servers and the search requirement of data servers, we should integrate the metadata and the data systems together to give users an integrated and consistent access to both data and metadata. Following this strategy, we have designed a Scientific Data and Information Super Server (SDISS) [Yang, Kafatos & Wang 2002] based on DIMES and GDS to tackle such problems. Combining a successful metadata interoperability solution such as DIMES with a data interoperability solution such as DODS will dramatically enhance the data accessibility.

In this paper, we first introduce the XML-based DIMES and briefly describe the DODS, in particular, the GDS system. Then, we describe the architecture of the SDISS and give a use case of the SDISS. Finally, we present the conclusions and future work for the SDISS improvement.

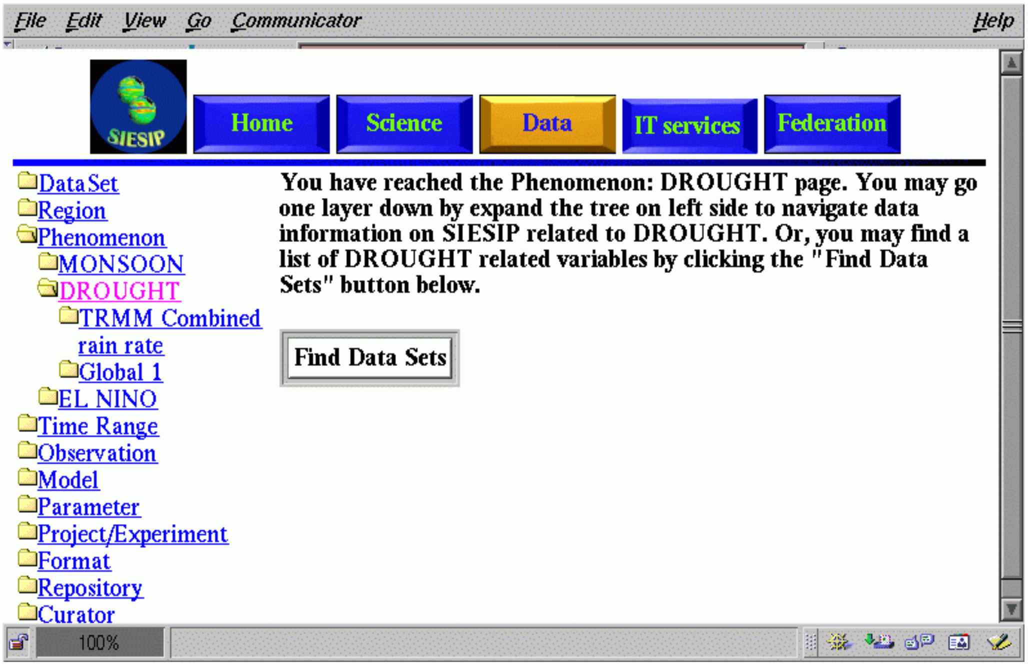
2. DISTRIBUTED METADATA SERVER (DIMES)

The Extensible Markup Language (XML) is an ideal standard to describe ASCII-based data, since data encoded in XML are understandable by both human users and computers. Most of the metadata for Earth science data are in ASCII format, and therefore can be easily migrated to XML format. We, therefore, developed the Distributed Metadata Server (DIMES) using XML technology. DIMES comprises a metadata model, an XML query engine as well as a web-based prototype interface.

Since metadata for Earth science data are varied in their contents and formats, and DIMES is designed for helping various data providers including those relatively small, a very flexible metadata model without much predefined schema is employed. A naive way to integrate metadata from heterogeneous sources is to represent metadata from different sources in XML format without additional restrictions, and pull together all these XML documents as the metadata repository. Though this approach requires minimum effort, it does not effectively support the search of metadata since there is no common agreement among these different XML documents. We, in DIMES, introduce some special attributes for the XML elements representing metadata concepts, and hence require some additional semantic enforcement. However, we follow the philosophy to minimize the semantics requirements: special attributes, if presented in the XML (metadata) documents, will be recognized and utilized automatically during the search of metadata. For XML documents without the special attributes, the tree structure in XML is used instead.

In the DIMES metadata model, a metadata concept is treated as a node. Special attributes are introduced to link nodes together, and metadata search can follow the links. The links reflect the logical relationship among nodes. Currently, in addition to the structural parent/child relation, we have defined the symmetric refer_to link, the paired node_types and type_instances links, and the inline_node link. One important flexibility feature of DIMES is that it allows data providers to define their own links to accommodate their special needs or understanding of the metadata. Another key feature of the DIMES framework is that the user-defined links can be utilized in the search through configuration of the DIMES search engine.

The DIMES search engine actually is a specific XML Query Engine, which is responsible for answering queries against the XML metadata. The main design goal of DIMES is flexibility, and the query engine is designed using the same strategy. Instead of defining fixed queries for specific purpose, we identify a few fundamental types of queries, and use these queries to answer complicated user requests. The first type of queries are basic queries which support regular queries in Earth science data systems such as finding data based on spatial/temporal coverage etc. Another type of query is the Nearest Neighbor Search, through which users can find closest nodes related to the current node. An important use and a special variation of the nearest neighbor search is the tree expand query. Essentially, the tree expand query is to re-organize a graph (nodes linked with different relations) into a tree with a specified root node. In practice, the tree expand query paves a path for DIMES client designers to develop client interfaces to present the metadata in a way that users can navigate easily and understand quickly.



You have reached the Phenomenon: DROUGHT page. You may go one layer down by expand the tree on left side to navigate data information on SIESIP related to DROUGHT. Or, you may find a list of DROUGHT related variables by clicking the "Find Data Sets" button below.

Figure 1. A screen shot for the metadata navigation prototype.

One common search interface for Earth science data systems is web forms supporting traditional search criteria such as spatial, temporal, and textual searches. DIMES supports such common search. In addition, one innovative interface is the metadata navigation system based on the special nearest neighbor search and the tree expand query. Figure 1 is a screen shot for the metadata navigation prototype. The advantage of this navigation system is that it allows users to navigate the metadata holding as browsing a local file system. Moreover, from any navigation point, users can leverage the nearest neighbor search to find data sets conceptually closest to the local node. An example of an actual use scenario will be described in section 4.

3. GrADS/DODS SERVER (GDS)

The Distributed Oceanographic Data System (DODS) [DODS 2002] is a robust, client-server data transport protocol based on the HTTP protocol. The DODS infrastructure provides a generic and flexible data model with the capability to distribute both digital data and the metadata that describe those data. With a DODS-enabled application program, a scientist can open a data set with a URL instead of a local filename. In other words, DODS enabled application programs can be considered as special web browsers. The special browsers receive data though the HTTP protocol and handle the data based on capabilities of the programs such as data analysis and data visualization. In contrast, the regular web browsers such as Netscape and Internet Explore will display the DODS data stream as numbers in ASCII. The simple design and ease of use of DODS has led to its widespread adoption to distribute online digital data by a great many Earth science data providers.

Recently, DODS has been enhanced by tightly integrating DODS and GrADS (the Grid Analysis and Display System) [Doty *et al.* 1997] to enable the application of the power of GrADS analysis over the Internet. This GrADS/DODS Server (GDS) [Doty *et al.* 2001] extends the DODS to include additional data formats, notably GRIB (Gridded Binary adopted by the World Meteorological Organization), and enables on-the-fly, server side data analysis and manipulation. Scientists who use GDS can retrieve data and metadata, and they can overcome the delays inherent in the relatively low bandwidth Internet by distilling data to its essence at the server.

4. A SCIENTIFIC DATA AND INFORMATION SUPER SERVER (SDISS)

DIMES and GDS support metadata search and data access, respectively. In order to give scientists a more powerful system moving towards an end-to-end system of data search, access and analysis, we integrate DIMES and GDS to form a Scientific Data and Information Super Server (SDISS). SDISS is termed a super server because it supports *interactive access to both metadata and data*. The power of the super server is reflected by the consistency between the data server and the metadata server and by the closely coupling between them. Our goal is to maintain the consistency between the data holdings in the data server and the metadata contents in the DIMES all the time. As a result, we can guarantee that a data set can be found through the metadata engine and all data sets found through the metadata engine are accessible through the data server.

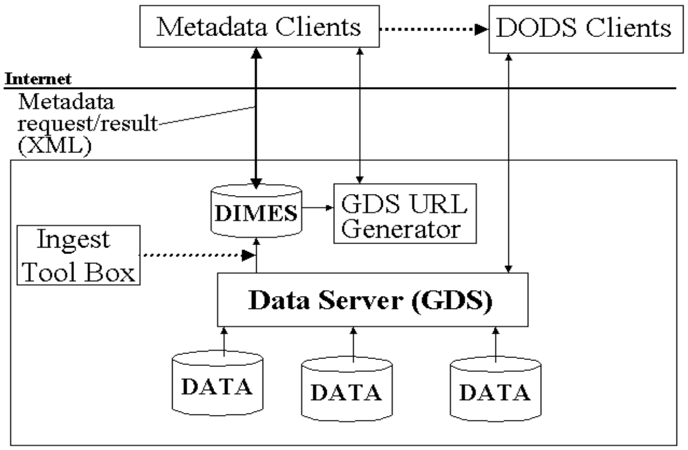
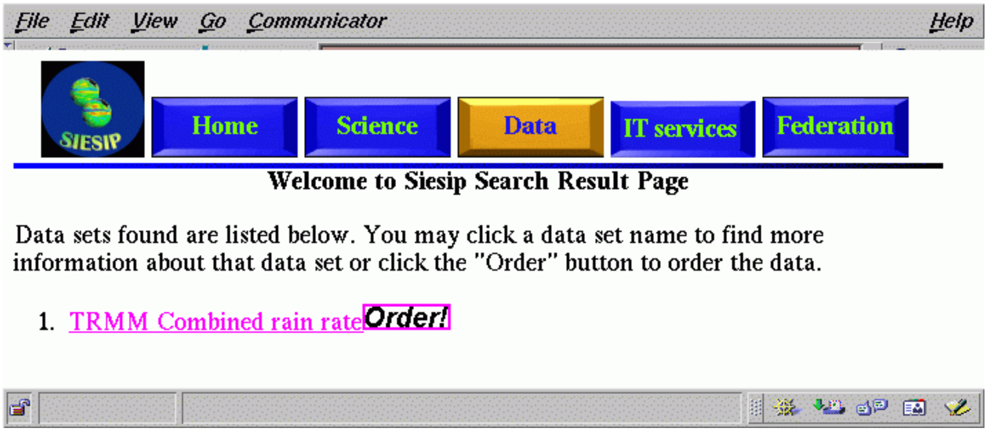


Figure 2. The high-level system architecture of SDISS.

Figure 2 is the high-level system architecture of SDISS. One component is not discussed above is the GDS URL generator. GDS allows users to fully leverage the power of GrADS to manipulate data on the server side before ordering the final products. The GDS URL's are relatively more complex than the plain DODS URL's. For GrADS users, it is straightforward to write a GDS URL. However, for non-GrADS DODS client users, the GDS URL's could be too complex to be used easily. To help such non-GrADS users, we developed a web-based GDS URL generator with predefined functionalities. The current GDS URL generator prototype includes only two functions, spatial average and temporal average. The URL generator, no matter how many functions will be supported, would provide only limited functionalities from those supported by GrADS. We expect that the generator will help GDS users to get familiar with the GDS URL's and then modify the sample URL's to create more specific URL's.

Another unmentioned item in the system architecture is the ingest toolbox. Metadata ingesting is very critical to maintain the consistency between the metadata server and the data server, especially in cases with routinely changed data server. There are many issues to confront the development of such ingest tool. We do not discuss the ingest tool here. Instead, we exploit a "use case" of such a super server to describe the usefulness of the system, and to discuss potential improvement.

A user would access an SDISS by first visiting a web page such as shown in Figure 1. The user could browse the metadata until find data sets of interest. The user could also search the system to find data sets from any browsing point. For example, the user intends to locate data sets based on phenomena. Therefore, the user browses the "Phenomenon" folder and then the "DROUGHT" phenomenon. From this browsing point, the user decides to search all related data sets. So the user click the "Find Data Sets" button provided in the interface. The search results are shown in Figure 3 with, in this particular case, only one data set having been found. This search mechanism can be utilized from any browsing points.



Welcome to Siesip Search Result Page

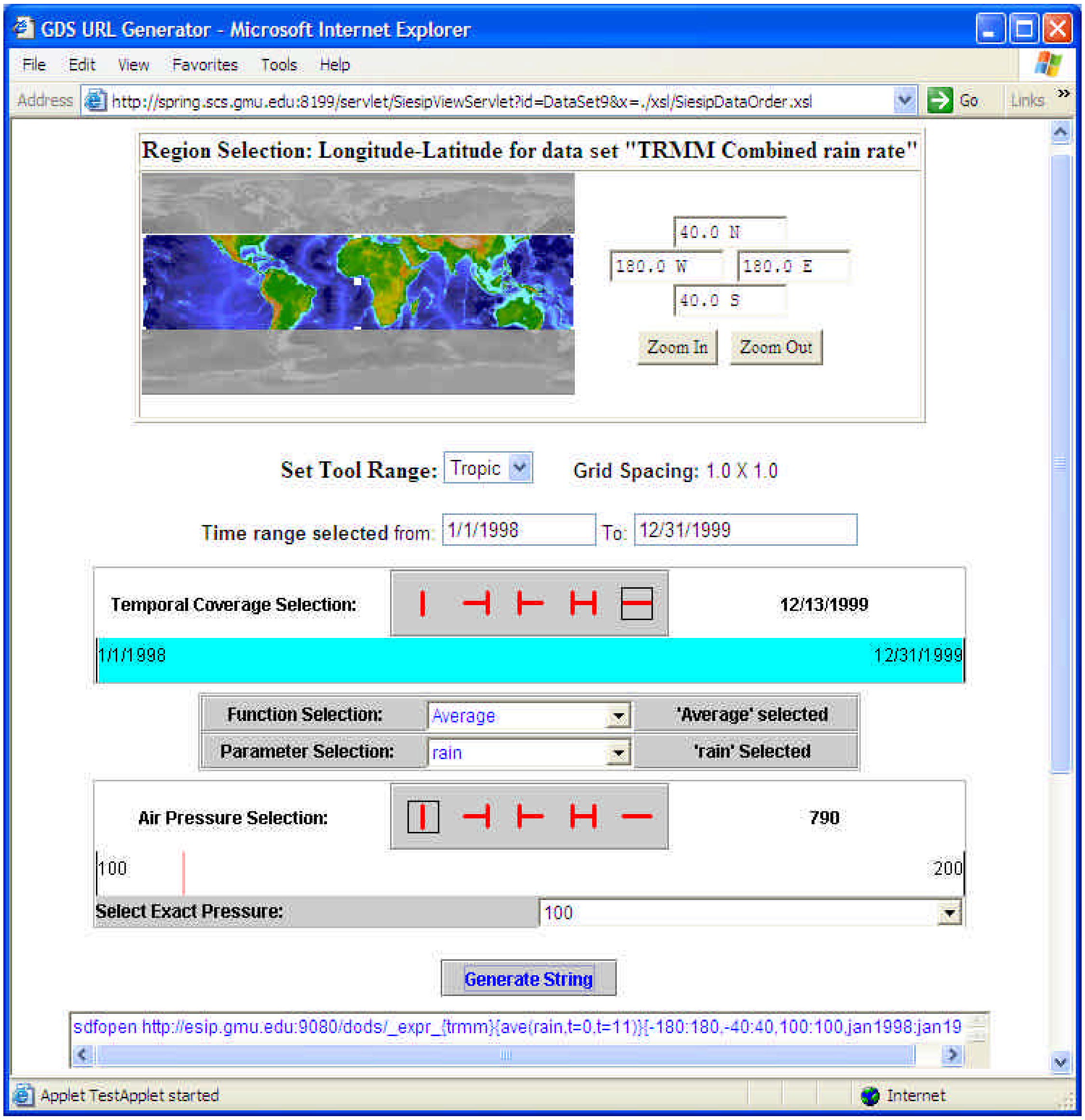
Data sets found are listed below. You may click a data set name to find more information about that data set or click the "Order" button to order the data.

1. TRMM Combined rain rate Order!

Figure 3. DIMES search results.

One further development, we have integrated DIMES with the web-based GDS URL generator and use the search result in DIMES to preload the metadata in the generator to better guide users to create their data requests. Broadly speaking, one may consider the result-driven GDS URL generator as a component of the SDISS.

One example of the GDS URL generator is shown in Figure 4. Actually, a user of the SDISS would get this interface by click the "Order" button shown in Figure 3 immediately after the data set name. Since the data set that the user wants to access is the TRMM (Tropical Rainfall Measuring Mission) rainfall data, and TRMM covers the tropical area only, the interface shows the valid area and allows the user to make selections in the valid area only. The spatial selection portion of the interface is build based on the LiveMap developed in NOAA/PMEL [Callahan 2002]. Similarly, the selection for a time or a time period will also be limited in the binding temporal coverage range. Moreover, the data set and parameter selection is also driven by the query result. DIMES allows two level data selections, data set level and parameter (variable) level. If a user orders data on the data set level, the system will preload the metadata about this data set and allow the user to select any parameter in this data set. If the user selects a particular variable before entering to the URL generator interface, this selected variable will be automatically used for generating the URL strings. Of course, the user can switch the variable before creating the URL even after entering the interface via a "variable order."



Address: http://spring.scs.gmu.edu:8199/servlet/SiesipViewServlet?id=DataSet9&x=./xsl/SiesipDataOrder.xsl

Region Selection: Longitude-Latitude for data set "TRMM Combined rain rate"

40.0 N

180.0 W 180.0 E

40.0 S

Set Tool Range: Tropic

Grid Spacing: 1.0 X 1.0

Time range selected from: 1/1/1998 To: 12/31/1999

Temporal Coverage Selection: 12/13/1999

1/1/1998 12/31/1999

Function Selection: Average 'Average' selected

Parameter Selection: rain 'rain' Selected

Air Pressure Selection: 790

100 200

Select Exact Pressure: 100

sdfopen http://esip.gmu.edu:9080/dods/_expr_{trmm}{ave(rain,t=0,t=11)}{-180:180,-40:40,100:100,jan1998:jan19

Applet TestApplet started

Figure 4. GDS URL generator interface.

After the user makes decisions on which area to focus, the temporal period, the geophysical variable and predefined function selections, the user can click the “Generate String” button in the interface. The GDS URL would be created based on user’s selection. In the example of Figure 4, we also include the GrADS command to open data through DODS protocol although it is not part of the URL. With most computers right now, the user can cut and paste the generated URL to his/her DODS client and handle the data sets opened in this way as any other data sets opened locally.

The above use case shows how easy it is to use the SDISS to find and access data. We have integrated metadata for search and metadata for use together and provide users a seamless way from the beginning to the end. It is expected that a scientist with basic Earth science knowledge does not need anything else such as downloading the data, reading data description documents, etc. before really using the remote data for products.

5. CONCLUSIONS AND FUTURE WORK

We have successfully built a Distributed Metadata Server (DIMES) and designed a Scientific Data and Information Super Servers (SDISS) by integrating the metadata server with an existing data server, GDS. The SDISS is the result of discussions with domain scientists and will be used to satisfy the Earth scientists’ needs for data search, data analysis, and data access.

Our next step is to develop and to integrate the ingest tool into the SDISS. Specifically, we need to develop a module to parse DODS metadata and to integrate the new “parser” to have an automatic DODS-to-DIMES metadata ingestion tool as shown by the dashed arrow line in Figure 2. The automation is expected to be difficult because the system need to handle the data adding/removing efficiently.

Another potential improvement on the SDISS is represented by the other dashed line in the system architecture (Figure 2). Currently, we assume the user agents for DIMES are regular web browsers. However, as we pointed out before, the data browsers for DODS data are specially tuned data analysis and visualization tools. To provide a uniform work environment for users to use SDISS, it is better to make DIMES directly available to some DODS-enable tools. That is, to develop an end-to-end, one-stop data information system serving Earth science communities.

ACKNOWLEDGMENTS

We acknowledge partial funding support from the NASA ESDIS ESTO prototype Project (NAG5-10177), from the Earth Science Enterprise WP-ESIP CAN Program (NCC-5306), as well as from George Mason University. The authors would also like to thank the following people for their valuable inputs: X. Deng and C. Wang of George Mason University for DIMES; B. Doty, J. Kinter, & J. Wielgosz of the Center for Ocean-Land-Atmosphere Studies for GDS.

REFERENCES

G. Asrar and R. Greenstone, eds., 1999: “EOS Reference Handbook,” NASA (Washington, D.C.)

Jonathan Callahan, 2002: “LiveMap version 3.0,” http://tmap.pmel.noaa.gov/~callahan/JAVA/map_v3.0/LiveMap_v3.0.html (last accessed on September 10, 2002).

DODS 2002: Distributed Oceanographic Data System. http://www.unidata.ucar.edu/packages/dods/ (last accessed on September 10, 2002).

B E. Doty, J. L. Kinter III, M. Fiorino, D. Hooper, R. Budich, K. Winger, U. Schulzweide, L. Calori, T. Hol, and K. Meier, 1997: “The Grid Analysis and Display System (GrADS): An update for 1997: ” 13th Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology, pages 356-358 (American Meteorological Society, Boston).

Brian E. Doty, Joseph Wielgosz, James Gallagher, and Daniel Holloway, 2001: “GrADS and DODS,” 17th International Conference on Interactive Information and

Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Jan 2001.

M. Kafatos, X. Wang, Z. Li, R. Yang, and D. Ziskin, 1998: "Information Technology Implementation for a Distributed Data System Serving Earth Scientists: Seasonal to Interannual ESIP," in Proceedings of the 10th International Conference on Scientific and Statistical Database Management (M. Rafanelli and M. Jarke, eds.), pages 210-215, IEEE, Computer Society.

R. Yang, X. Deng, M. Kafatos, C. Wang and X. Wang, 2001: "An XML-Based Distributed Metadata Server (DIMES) Supporting Earth Science Metadata," in Proceedings of the 13th International Conference on Scientific and Statistical Database Management (L. Kerschberg and M. Kafatos, eds.), pages 251-256, IEEE, Computer Society.

R. Yang, M. Kafatos, and X. Wang, 2002: "Managing Scientific Metadata Using XML," *IEEE Internet Computing,* v6, no.4, pp. 52-59.