9.7 AUTOMATIC METADATA INGESTION FOR SUPPORTING A WEB-BASED SCIENTIFIC DATA AND INFORMATION SUPER SERVER

Ruixin Yang*, Yujie Zhao, Menas Kafatos
Center for Earth Observing & Space Research (CEOSR)
George Mason University (GMU)

ABSTRACT

A DIstributed MEtadata Server (DIMES) system has been developed for representing, storing, retrieving and interoperating metadata in a distributed environment. DIMES supports both regular metadata search and a web-based metadata navigation mechanism. DIMES has also been combined with an existing data access and analysis server, namely the GrADS/DODS server, for the purpose of forming a Scientific Data Information Super Server (SDISS), which supports both metadata and data.

The SDISS guarantees consistency between the content of the data server and the content of the metadata server. The key element to achieve this goal is to ingest metadata from the data server to the metadata server. In this paper, the major issues in the ingestion procedures, the design of the ingestion system, and its implementation are discussed and ingestion examples are given.

1. INTRODUCTION

As more and more data for Earth science research are available from Earth observing, in particular, satellite remote sensing, and from models, efficient mechanisms for finding and delivering distributed data become necessary. One well-known data delivery infrastructure is the Distributed Oceanographic Data System (DODS) originating in the oceanography community (DODS, 2003). Plain DODS was enhanced by being combined with the data analysis capability of GrADS (the Grid Analysis and Display System) (Doty *et al.,* 1997) to form GDS (GrADS/DODS Server) (Wielgosz *et al.,* 2001), which allows users to define operations performed on the server side and to obtain the resultant information (processed data) via the Internet.

There exist many metadata systems serving the Earth science community for finding Earth science data. A major problem with the current search systems, usually large and supported by national centers, is that there are too many hits for a specific search and some of the data links may be out of date. One solution to this general problem is to have a metadata server consistent with a data server all the time. Following this strategy, a Scientific Data and Information Super Server (**SDISS**) (Yang, Kafatos & Wang, 2002; Yang *et al.*, 2003) was designed by enhancing GDS with a metadata server, i.e., DIMES (DIstributed MEtadata Server).

The XML-based DIMES (Yang *et al.,* 2001) implemented a flexible metadata model and web-based metadata navigation interfaces to support various level metadata accesses. In DIMES, a metadata concept is treated as a node in the XML DOM (Document Object Model) model. It is expected that combining a successful metadata interoperability solution such as DIMES with a data interoperability solution such as GDS will dramatically enhance the data accessibility for Earth scientists.

2. SDISS ARCHITECTURE AND MAJOR COMPONENTS

Figure 1 is the high-level system architecture of SDISS. Major components of the system are certainly GDS and DIMES. A GDS URL Generator is included in the architecture to help users to build the relatively complex GDS URLs through a GUI. The SDISS is designed

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

to be a distributed system, and therefore a register can be used to record all available SIDSS. Certainly, the SDISS register itself can be distributed providing information on each server or even residing on the client side. However, the centralized register will make it easier to reach broader audiences by leveraging the existing centralized metadata search engines such as GCMD (Olsen & Major, 1996). On the server side, setting up a SDISS starts both the data server and the metadata server. Metadata in an SDISS will be ingested and checked to reflect the changes on the data holdings served by the current data server. Since the data server may be updated by adding data to or removing data from the server, the major challenge is to keep the data and the metadata consistent in an existing SDISS. A ingest tool box is created for achieving this goal, and the ingestion process and corresponding tools are the focus of this article.

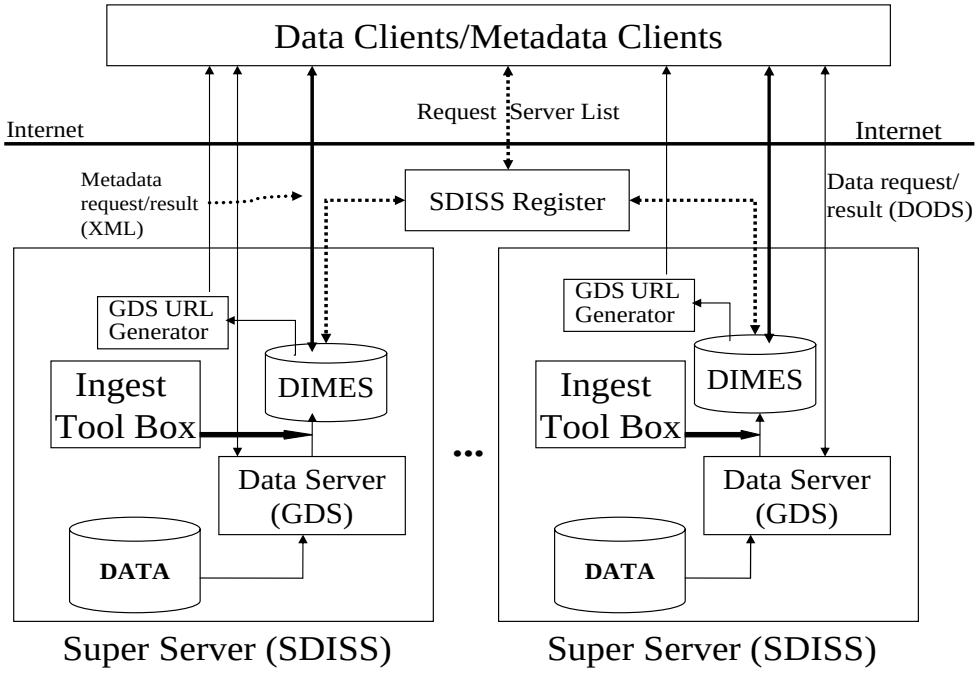

Figure 1. The high-level system architecture of SDISS.

3. METADATA INGESTION

Figure 2 is a flow chart of the SDISS start up and metadata ingestion procedure. A data provider wants to set up an SDISS by editing a configuration file that defines data sets supported by the GDS. The GDS setup program will check the original metadata and ingest the metadata into GDS metadata. To make the GDS metadata searchable, the GDS metadata are ingested into DIMES metadata. Two level ingestion/cleaning procedures are used for the full DIMES metadata setup process. The first level consists of a data set list. By comparing the current data sets supported by GDS with those in the previous GDS, we can quickly have an add/delete list, which records data sets added and deleted. Then, the DIMES ingester ingests all metadata into a temporary metadata in DIMES format. However, we cannot use the new metadata directly. Instead, we need to merge and clean the new metadata with the existing copy. The ingestion/cleaning/merging tasks are finished in the level-two procedure. At this level, first, we pick the existing DIMES metadata as the current metadata, and then ingest the information in the add/delete list by deleting nodes corresponding to the data sets no longer in service and adding new nodes from the temporary metadata directly into the current DIMES. The difficult part is to handle datasets already in service. In Earth science,

new data may be added to existing data sets as time goes on. The data sets keep all features and names except for the temporal coverage. A more special case is for operational data servers in which not only new data sets are added but also some old data sets need to be deleted, in other words, treating the process as if we have a data pipe in service. The ingest tool box for the level-two procedure has been developed with considerations of the special “data pipe” case.

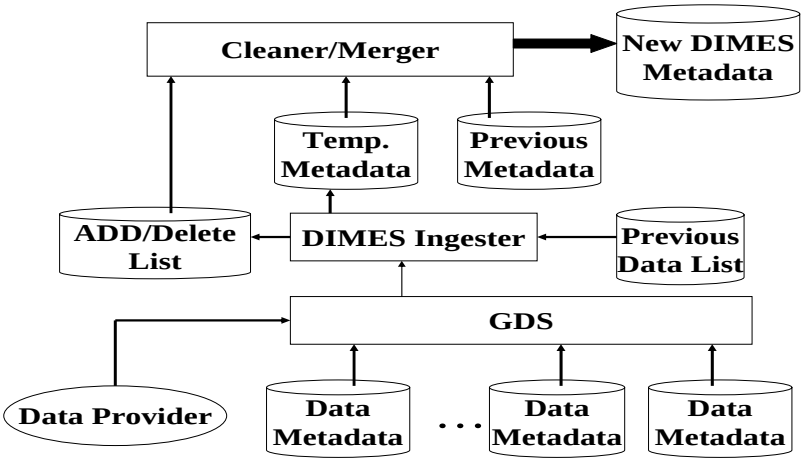

Figure 2. Metadata Ingest Flow Chart.

The tool box includes software components for both metadata-to-XML ingestion and XML-to-XML merging and cleaning. The ingestion software is coded in Java because of the flexibility of Java-XML DOM technology and for a consistency with our XML query engine written in Java. Since the near term goal is to integrate DIMES into GDS, the ingestion programs harvest metadata directly from a GDS server. Nonetheless, the software can be easily extended to ingest other kinds of metadata since the output parts for all cases are the same.

Once the GDS metadata are converted and ingested into DIMES format, the metadata need to be merged with existing metadata repositories and to be cleaned. DIMES allows domain knowledge and user annotations in the metadata to be added. Such knowledge is usually not in the original metadata and is accumulated over time. Therefore, to reserve and use such knowledge, one cannot simply replace DIMES metadata with the new one after each SDISS update. Instead, one should merge the new harvested content into the current metadata. It is also necessary to clean the new merged metadata to maintain consistency and efficiency of the overall metadata repository.

During the cleaning process, redundant nodes are deleted. In addition, some data sets grow over time, and the ingestion tool might consider new data from the same sets as independent data sets and create separate metadata entities for them. In this case, we need to keep only one copy of the metadata, but extend the temporal coverage to describe both old and new data. Since DIMES uses linked nodes as the basic metadata model, and some links appear in pair and some of them are symmetric, the cleaner programs check these properties and make necessary addition and modifications.

The cleaning and merging programs are well separated from the ingestion programs since metadata from other sources (not ingested) can also be checked using this tool. Moreover, the XSLT technique is used to define duplications among nodes, which separate the domain knowledge from the programming logic of the tool.

After both level one and level two procedures, a new DIMES metadata will be ready for use. The metadata retain all the knowledge

developed before and integrate metadata of new data sets into the online SDISS. The current metadata will be consistent with the current data server, and therefore avoid the common non-valid link problem. It will also guarantee all data sets are searchable so that users will not miss the data sets they are looking for. Of course, knowledge about the new added data sets cannot be automatically added at the present development stage. In the metadata ingestion process, intelligent methods could be introduced to capture additional knowledge not in the original metadata.

4. CONCLUSIONS AND FUTURE WORK

We have successfully built a Distributed Metadata Server (DIMES) and designed a Scientific Data and Information Super Server (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.

The major programs for the ingest tool box have been developed and tested again a GDS at George Mason University (http://spring.scs.gmu.edu:9080/) and another GDS at National Centers for Environmental Prediction (NCEP) (http://nomad2.ncep.noaa.gov:9090/dods). Metadata for those servers are well resolved in the harvested DIMES. However, much work needs to be done for achieving good performance for a large GDS server. Although significant progress has been made in the automatic metadata ingestion and system integration, the DIMES and its corresponding ingest tool box have not been integrated into a GDS yet.

ACKNOWLEDGMENTS

We acknowledge partial funding support 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: S. Dasgupta, Y. Nie, and X. Wang of George Mason University for DIMES; B. Doty, J. Kinter, & J. Wielgosz of the Center for Ocean-Land-Atmosphere Studies for GDS.

REFERENCES

DODS 2002: Distributed Oceanographic Data System. http://www.unidata.ucar.edu/packages/dods/ (last accessed on October 17, 2003).

Doty, B. E., 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).

Olsen, L.M., G. R. Major, "Global Change Master Directory Enhances Search for Earth Science Data," 1996, Transactions of the EOS, American Geophysical Union, http://www.agu.org/eos_elec/95127e.html (last accessed on October 17, 2003).

Wielgosz, J., B. E. Doty, J. Gallagher, and D. Holloway, 2001: "GrADS and DODS," 17th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, Jan. 2001.

Yang, R., 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.

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

Yang, R., X. Wang, Y. Nie, Y. Zhao and M. Kafatos, 2003: "A Web-Based Scientific Data and Information Super Server with A Flexible XML Metadata Support," in Proceedings of the 19th International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology, American Meteorological Society, *Long Beach, California, February 9-13, 2003 (CD-ROM).*