P1.1

AN INTEGRATED DATA MANAGEMENT, RETRIEVAL AND VISUALIZATION SYSTEM FOR EARTH SCIENCE DATASETS

Zhenping Liu *, Yao Liang *
Virginia Polytechnic Institute and State University

Xu Liang **
University of California, Berkeley

ABSTRACT: This paper describes our work in developing an integrated data management, retrieval and visualization system for earth science datasets with extensibility, scalability, uniformity, transparency and heterogeneity. XML based metadata mechanism is the foundation of data management in our system. Dynamically generated query GUI makes it easy and convenient for scientists to access and retrieval diverse datasets. Scientific visualization toolkits display huge amount of data graphically to help researchers have better understanding of the data and gain valuable insights of the datasets under investigation. This system helps earth scientists use, share and visualize data more efficiently. Without knowing any information of the physical storage location, content, structure and format of each dataset instance, and without programming a single line of codes, scientists can now query heterogeneous data easily, and view and understand the retrieved data in analytical and graphical ways.

1. INTRODUCTION

The community of earth science requires processing and analyzing not only a large amount but also a variety of data from diverse distributed sources, such as space and ground-based observations. These data are usually stored in many different data formats associated with different data management systems. Due to the wide variety of data formats and structures, scientists have to spend a significant amount of effort to convert the data from their original heterogeneous formats and structures into usable information: such as writing specialized computer front-end programs – one for each kind of datasets before using these data to conduct any scientific analyses [1].

Another challenge that earth scientists often face is the difficulty in effectively exploring and visualizing a huge amount of data, especially the multi-dimensional data, which could then significantly deteriorate researchers' efforts from gaining meaningful knowledge out of the large amounts of data under investigation. However, most data publishers do not yet provide these services at present. On the other hand, it is difficult and very time-consuming for earth scientists to develop advanced data analyses and visualization applications by themselves alone.

In order to help earth scientists to use, share and visualize heterogeneous data more efficiently so that they do not need to spend much time working as computer programmers but focus on the research in their areas, we develop an integrated data management, retrieval and visualization system, which works like a middleman between the scientists and the diverse data sources, to coherently manage, share and analyze data volumes with extensibility, scalability, uniformity, transparency and heterogeneity. The system provides a metadata-based integrated environment, which allows scientists to access all of the datasets in the system by specifying query conditions, to customize the format of any retrieved data files as they need, and to browse, conduct statistical analysis, and visualize the retrieved data online. All of these operations are done through a uniform GUI (Graphical User Interface) generated dynamically according to the metadata of dataset instances. With our system, scientists can obtain data at their request, view and understand the retrieved data in analytical and graphical ways without need to know any

* *Authors address:* Zhenping Liu, Yao Liang, Alexandria Research Institute, Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, 206 N. Washington Street, Suite 400, Alexandria, VA 22314; Email: {zhliu, yaliang}@vt.edu.

** *Corresponding author address:* Xu Liang, Department of Civil and Environment Engineering, University of California, Berkeley, CA 94720; Email: liang@ce.berkeley.edu.

information about the physical storage, original structure and format of dataset instances, and without need to programming a single line of codes.

2. DATA MANAGEMENT AND XML BASED METADATA

2.1 *Metadata – Solving the Data Management Problem*

The complexity of managing and accessing a large amount of diverse datasets is becoming one of the most challenging problems for users of large earth data systems. How can we manage and access data intelligently and efficiently, given such increasingly complexity? One good solution would be to create the so-called metadata for recording physical and conceptual details about the data under management. These metadata provide more system-level information than standard file storage systems and enable intelligent and efficient access and management of the data. Metadata make it possible for different applications (using metadata and metadata management systems) to share data through common interfaces and structures.

In our system, metadata mechanism is at the heart of data management. It provides a dynamic, extendable, scalable and efficient architecture to manage disparate data sources. It eases the process of data, makes it possible to organize, share, search, retrieval and view data. Specifically, an XML (Extensible Markup Language) [2] based model for the metadata is developed and implemented in our system to describe the structure, storage, category and content information of any earth science datasets in any data formats.

2.2 *XML Based Metadata*

Metadata here contain two parts. XML file, and XML schema. XML files are written using a number of elements and attributes to describe various datasets. There are four types of XML files created with different functions, which are (1) To describe category of data files by data sources or by physical data types, (2) To describe storage information of data files, (3) To describe content and semantic of data files and allow system to understand the meaning of the data, and (4) To describe format and organization of data files. XML schema defines the elements and attributes used in XML files, and describes the structure of an XML document and specifies legal building blocks of an XML file, which can be used to validate an XML file.

2.3 *Managing Metadata with XML Editor*

As stated, the key to manage data is to manage their metadata. However, for scientists not familiar with XML, writing XML metadata for new data source or managing existed XML files can be potentially complicated. However, there are many free [e.g., 3, 4, 5, 6] or commercial [e.g., 7, 8, 9, 10] XML editors, which provide user-friendly GUIs for scientists to specify the structure, storage, category and content of the earth science data files without knowing much of the XML schema and rules. The editor makes it convenient for users to manage metadata.

2.4 *Flexibility and Extensibility*

By employing the XML metadata technology, our system is so designed that it is flexible to be extended or changed with any new needs and any incremental development of the system.

(1) Adding New Data Sources

The user can add new data sources easily by simply capturing information about the new data source and creating new metadata files to describe it with an XML editor. There is no need to change the system.

(2) Creating New Applications

Based on XML metadata, we have implemented data management, search and retrieval (see section 3), and some applications. Currently, applications we have implemented include data viewer, analytical functions and scientific visualization methods. In the future, we will develop relevant data mining, fusion and some other analysis applications based on metadata. Scientists may also utilize existed XML metadata to develop their own analysis toolkits with our system.

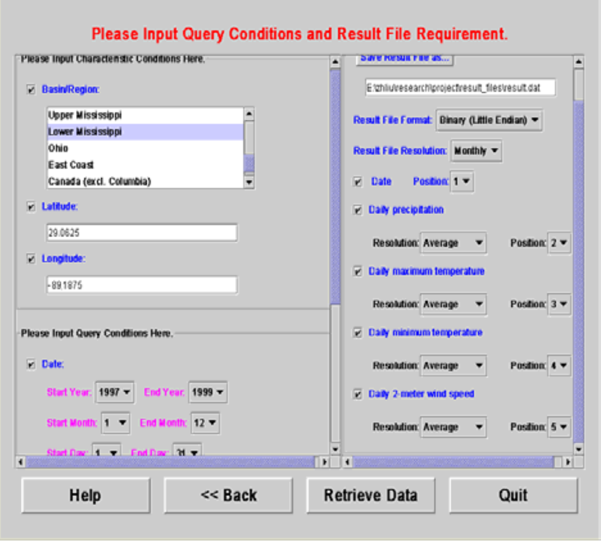
(3) Dynamic Applications Enabling

Since there are a variety of datasets in the system, apparently, not every application is suitable for all kinds of datasets. For example, for a time-series featured dataset, 2-D (two-dimensional) plots may be applicable while 3-D (three-dimensional) visualization may be not. How do users know which applications could be employed to a specific dataset? To achieve our goal of transparency to users, we have built metadata to describe corresponding relationships between dataset models and available applications. Based on these metadata, the system will configurate GUI dynamically for the dataset retrieved, so that only those relevant applications will be showed on the GUI, which facilitates users to pick up suitable applications provided conveniently when analyzing the retrieved data online.

3. DATA QUERY AND RETRIEVAL

3.1 *Dynamically Generated Query GUIs*

To allow users to be able to query data system without background knowledge and training, we provide query GUIs for all datasets in the system. Our approach is to create a data query system that dynamically creates dataset query GUI for diverse datasets based on characteristic of diverse datasets. These characteristics are described in XML metadata. By using stored metadata to create the query interfaces, a standardized yet dynamic system is created that allows querying of assorted datasets. By this way, the system eliminates the need to create custom programs for different datasets. When adding new datasets to the system, query GUI for these datasets will be dynamically generated if the characteristic of these datasets has been specified in metadata. Therefore, the system provides an extensible and scalable query GUI framework. Through dynamically generated query GUI, scientists can specify search conditions, customize the format and the resolution type of the result files as needed.



Please Input Query Conditions and Result File Requirement.

Please Input Characteristic Conditions Here.

☑ Basin/Region:

Upper Mississippi
Lower Mississippi
Ohio
East Coast
Canada (excl. Columbia)

☑ Latitude: 29.0625

☑ Longitude: -89.1875

Please Input Query Conditions Here.

☑ Date:

Start Year: 1997 End Year: 1999

Start Month: 1 End Month: 12

Start Day: 1 End Day: 31

Save Result File as...

E:\zhliu\research\project\result_files\result.dat

Result File Format: Binary (Little Endian)

Result File Resolution: Monthly

☑ Date Position: 1

☑ Daily precipitation

Resolution: Average Position: 2

☑ Daily maximum temperature

Resolution: Average Position: 3

☑ Daily minimum temperature

Resolution: Average Position: 4

☑ Daily 2-meter wind speed

Resolution: Average Position: 5

Figure 1 A query GUI sample dynamically generated

3.2 *Query Categories*

To make it convenient, users are able to search for data by the category of physical data types or data sources. Users may also define new categories by adding their definitions into metadata and the system will then dynamically add them into GUIs.

3.3 *Data Search and Retrieval*

The data retrieval is based on the metadata system, that is, the description of the physical datasets. After user submit query request through GUI, the query can be then devised to search over the metadata catalogue, which is implemented in hierarchy. Once all data files that possibly satisfy the query have been identified by searching the available metadata, the system will retrieve these files holding the actual data and obtain useful result data. If the resolution type is not the same as original request, it will do computation and obtain data in new resolution. Then it organizes the result data file in the format that users specified.

4. SCIENTIFIC VISUALIZATION AND ANALYSIS FUNCTIONALITIES

Scientific visualization is concerned with the interactive display and analysis of data. It facilitates researchers to gain valuable insights of the data under investigation. It can reveal correlations between different quantities both in space and time, and it opens up the possibility to view the data selectively and interactively online. Therefore, it can assist researchers in extracting the hidden information of the data, whether empirical data gathered from observers/sensors or derived data generated by computational models through the visualization of those original data [11]. In our system, we have developed several scientific visualization applications to support the following visualization methods.

4.1 *2-D Plots*

2-D plots method displays the evolution of data along time. It also provides the basic statistics and zoom-in/out tools to allow users to interactively view data and calculate statistics (including averages, variance, maximum, and minimum) of the data within specified temporal range online. Figure 2 illustrates a 2-D plot of retrieved data.

4.2 *2-D Colormaps, 2-D Contours, and 3-D Surface Rendering*

2-D Colormaps represent 2-D scalar fields by mapping data values into colors. 2-D contours represent 2-D scalar fields by drawing curves representing constant function values. In 3-D surface rendering, depicts a function of the form $z = f(x,y)$ by a surface in space. The z values can be mapped into colors on the surface. Examples of 2-D colormap, 2-D contour and 3-D surface rendering of the NEXRAD Stage III Daily Precipitation for the Blue River Basin region on May 6 1995 are illustrated in Figures 3, 4, and 5 respectively.

4.3 *Animation*

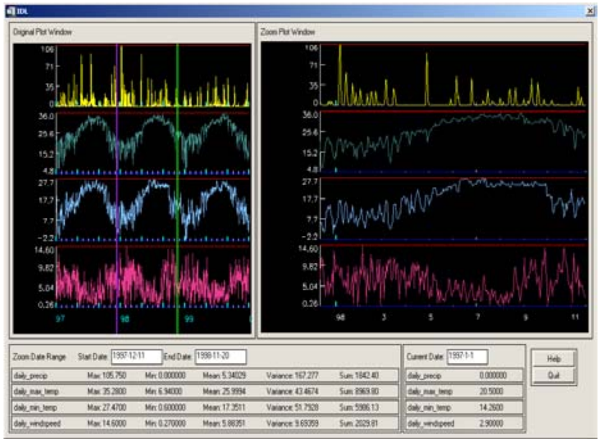
Animation provides flow control for time series data. It produces a movie of sequential time frames and shows the evolution of the data over time. The 2-D Colormaps, 2-D Contours, 3-D surface rendering pictures can be animated in our system.

4.4 *Data Probe (1-D, 2-D, 3-D domain)*

Data Probe is interactive querying of visual display at a specific point to obtain numeric value(s) at that point.

4.5 *Output Graphics Results to Files and Printer*

The system can export graphics results of data visualization to following format: Bitmap File (BMP, JPEG, TIFF, PNG, PPM, SRF), Bitmap Postscript, Vector Postscript, VRML, and can also export graphics results to a printer.



Zoom Date Range Start Date: 1997-12-11 End Date: 1998-11-20

daily_precip	Max: 105.750	Min: 0.000000	Mean: 5.34029	Variance: 167.277	Sum: 1842.40
daily_max_temp	Max: 35.2800	Min: 6.94000	Mean: 25.9994	Variance: 43.4674	Sum: 8969.80
daily_min_temp	Max: 27.4700	Min: 0.600000	Mean: 17.3511	Variance: 51.7928	Sum: 5986.13
daily_windspeed	Max: 14.6000	Min: 0.270000	Mean: 5.88351	Variance: 9.69359	Sum: 2029.81

Current Date: 1997-1-1

daily_precip	0.000000
daily_max_temp	20.5000
daily_min_temp	14.2600
daily_windspeed	2.90000

Figure 2 2-D plot

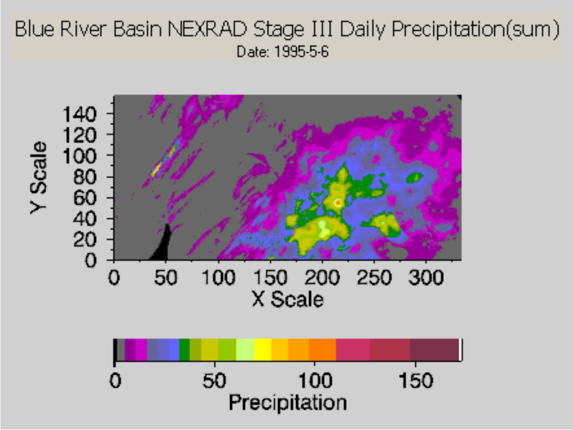


Figure 3 2-D colormap

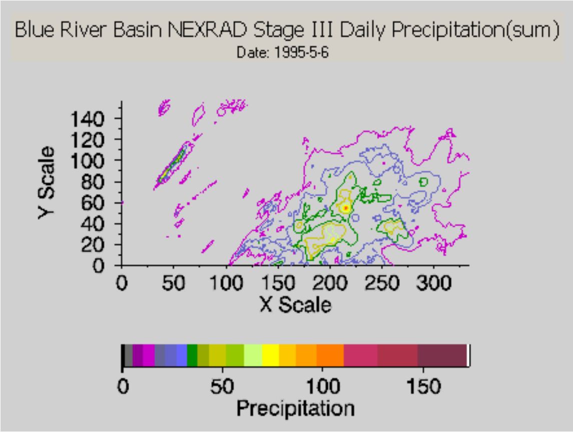


Figure 4 2-D contour

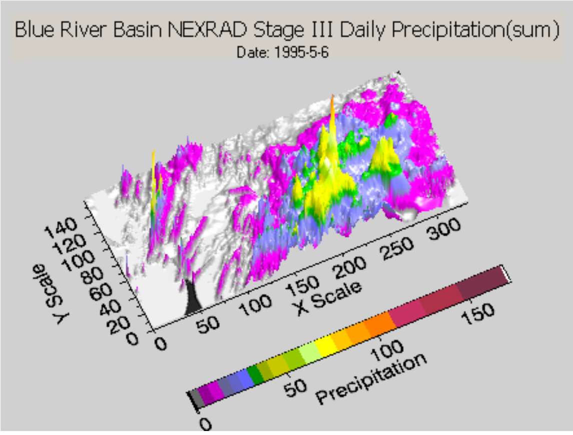


Figure 5 3-D surface

5. SUMMARY AND FUTURE PLANS

The data management, retrieval and visualization system for earth science datasets presented in this paper provides an extendable, scalable, uniform and transparent environment to use, share and visualize diverse earth science datasets. It allows earth scientists to spend much less effort/time on programming and data handling, but rather to concentrate on their own research areas.

Currently, the entire system is a standalone data service system. In the near future, we plan to employ the state-of-the-art Grid Service technology, Open Grid Services Architecture (OGSA) [12], to develop middle-ware infrastructure to coherently manage and share petabyte-scaled earth science data volumes in grid environments. OGSA is a web-services-based grid architecture comprising a set of interfaces and their associated behaviors to facilitate distributed resource sharing in heterogeneous multi-institutional dynamic environments. We also plan to develop more additional advanced analysis applications based on metadata to enhance the system's functionalities, such as data mining, data fusion and other advanced scientific visualization applications (such as map projections, virtual reality, and other 3-D visualization).

At present, the system supports binary (big endian and little endian) and text file formats, and will be extended to support CDF, HDF, HDF-EOS, netCDF in the near future. The system currently can perform time resolution transformation, and will be extended to process space resolution transformation.

REFERENCES

[1] Earth Science Markup Language (ESML), http://esml.itsc.uah.edu/

[2] XML: Extensible Markup Language, http://www.w3.org/XML

[3] XED, http://www.ltg.ed.ac.uk/~ht/xed.html

[4] MERLOT, http://www.merlotxml.org/

[5] LOGILAB's XML Editor, www.logilab.org/xmltools/xmleditor.html

[6] VISUAL XML, http://www.pierlou.com/visxml/

[7] Xmetal, http://www.softquad.com/top_frame.sq

[8] XMLwriter, http://www.xmlwriter.com/

[9] Morphon XML-Editor, http://www.morphon.com/xmleditor/index.shtml

[10] XML Spy Document Editor, http://www.xmlspy.com/download.html

[11] Scientific Visualization Laboratory, http://www.scivis.gatech.edu/

[12] The Globus Project, http://www.globus.org/