

New Paradigm For Search and Order in EOSDIS

Robin Pfister

NASA/GSFC

Code 423, NASA/GSFC, Greenbelt, MD 20771

Phone: (301) 614-5171/Fax: (301) 614-5267/e-mail: robin.pfister@gssc.nasa.gov

Keith Wichmann

Global Science & Technology, Inc.

6411 Ivy Lane, Suite 300, Greenbelt, MD 20770

Phone: (301)474-9696/Fax: (301)474-5970/e-mail: wichmann@gst.com

INTRODUCTION

Through the 1980's and 1990's, iterative search and retrieval was the defacto data access paradigm. For the first 10+ years of planning the EOSDIS system, this search and retrieval paradigm was accepted by managers and systems designers as the system interaction standard. Since very few data and information systems existed at the time, users accepted this paradigm and worked within its framework. In the past year, another data access paradigm has emerged and is gaining popularity among members of the EOSDIS user community. The new approach is one of discovery and navigation where the user interface shows the user everything that is available and the user simply navigates to what is needed. We have also seen that some of our more serious or routine users have special needs and require very flexible user interfaces to meet their special needs. To meet these new challenges, EOSDIS is developing a new information management system that will support varied paradigms needed by our user community. This approach calls for an architecture that requires different protocols and standards than previously used.

BACKGROUND

EOSDIS is designed as a one-stop-shop for earth science data. Archives all over the world that accept and follow a few common protocols and standards are connected through a common search and order interface. The idea is to save duplication of effort by providing a single common interface to distributed data. This was based on the assumption that scientists have the same general needs when it comes to

search of and access to data. We are now finding that one size does not fit all. Each science group has special needs for their special datasets.

CHALLENGE

In the EOSDIS design, the search and retrieval paradigm depends on a small volume of common metadata to guide users through search construction. Although the original intent was to reduce the chance of nonsensical queries, it is still possible to issue a query that results in zero hits. This can be frustrating to users who can spend a significant amount of time constructing the right query criteria only to find no data. Frustrated users would comment "Just tell me what you have, then I'll tell you what I want". Rather than performing time consuming queries, users want the system to present the available data so they can simply navigate to desired data. To support the new paradigm, the user interface needs immediate access to large amounts of metadata. This has been illustrated in prototype systems such as Dynamic Queries, developed by the University of Maryland Human-Computer Interaction Laboratory [1].

The current EOSDIS configuration does not support this new paradigm but can be easily modified to support the population of an external metadata clearinghouse that will meet this need. One concern paramount to the success of such a system is ease of participation by data providers. The system must make it easy for data providers to provide their metadata to the clearinghouse or they are unlikely to share their metadata.

SOLUTION

The Independent Information Management System (IIMS) is being developed as a metadata clearinghouse that will provide flexibility for end users while keeping provider participation as simple and easy as possible. The use of the eXtensible Markup Language (XML) enables the architecture to meet these needs.

The IIMS approach is to gather as much remote sensing metadata as possible in one location known as the clearinghouse. This models e-commerce web sites where a catalog of available, orderable items is maintained at one location even though the actual items may come from multiple providers. As long as a client has network connectivity to the clearinghouse, it is possible to search the entire set of data. Also, hardware resources can be concentrated at this single point to insure constant operation using well-known load-balancing techniques. Combined, these architectural features allow for a system that responds very quickly. Quick performance is necessary to support new data access paradigms.

LAYERED ARCHITECTURE

The IIMS addresses extensibility in its basic architecture through its ability to expose a consistent, well-conceived interface that external systems can communicate with. Extensibility is achieved through a layered architecture based on Enterprise Java Beans (EJBs). This approach yields a system that can support interfaces of various forms. The IIMS also captures results of each step in finding and accessing data in order to maintain "context". New modules can be introduced that transform one context to another. The user of the system can then use the set of modules that works most like they require, freely interchanging modules that transform the same contexts. Based on user selections, different modules are chained together behind the scenes, passing context from one to the next, so the user experiences a seamless path through the system. Any module could be replaced in the system as long as it properly handled its input and output contexts. A new data search and access paradigm can easily be introduced by adding new steps in the process or simply changing the way modules work.

In cases where a change in paradigm cannot be accommodated by clever applications of existing query capabilities, an EJB can be constructed that has a closer relationship with the system. This bean can be used to reflect the desired data in the system, then any module can communicate with it. As an example, the discovery and navigation paradigm might employ a user interface to display the most recent data gathered from a satellite. The new module would need a simple way to get the pertinent data, and enough associated metadata to display it to the user appropriately. This display could be an interlinked map and dateline. A click on the displayed data on the map could display a browse image or even go directly to an ordering screen. Obviously this type of interface is more specialized, but the modular architecture of the IIMS supports its use simultaneously with the existing systems.

XML USAGE

Throughout the IIMS system, XML is used as a standard basis of storage and exchange. XML is a standard method of expressing information in a structured way through the use of tags. As a standard it only describes how tagged information is placed into parseable strings and exerts no constraints on the content of those data structures beyond defining the structure of the data for each message. Data Type Definitions (DTDs) are used to describe the structure of the data in much the same way that a template works. The DTD is used to check that an XML string is well-formed but does not check the contents of the structure.

In the IIMS messages passed within the system and persistent entities used to represent items such as contexts, are represented in XML. Queries and results are streams of XML data. The IIMS development team is making every effort to include the growing number of standards that are extending XML. For instance, the Geospatial Markup Language is intended to address using XML to describe geospatial data in a standard fashion. XQL is a standard query language expressed using XML. Earth Science Markup Language (ESML) is striving to establish a standard XML-based lexicon for describing Earth Science data. The team is using these efforts and others as a basis for defining the metadata interchange mechanism, the context description language and the data manipulation language. The use of

XML means that a number of tools can be applied to the system without having to develop new interfaces, or completely new compatible tools. Currently XML parsers and DTD validators exist to validate and translate commands into something that a system's business logic can manipulate effectively.

The IIMS will expose the clearinghouse metadata to internet search systems via XML. Current internet search engines rely on using the entire contents of a web page to drive a text search engine without any concept of what the text is describing. For example, a search for the words bill drafter might return documents about jobs for people who write legislation, documents written by Bill Drafter, and the actual legislation. XML allows documents to be stored with a tagged structure. If the lexicon of the tags is known, then a more intelligent search can be performed. The above search could be focused by specifying that only CONTENT of type JOB LISTINGS be searched for jobs for bill drafters. This search will return a more accurate set of results as long as the participants in the search subscribe to the same lexicon of tags. Many databases that are available on the web today provide their own interfaces for supporting this kind of search, but there is no generic coordinated method for searching multiple databases simultaneously. XML with a standard lexicon provides a step towards making this kind of search available through standard internet search portals.

SUPPORTING MULTIPLE DATA PROVIDERS

The IIMS targets the EOSDIS Core System (ECS) as its initial provider but it is intended to support multiple providers of a variety of types. The use of XML as the metadata description language allows existing tools to extract data out of commercial databases that are used for storing the provider's metadata. The IIMS team is developing a tool that allows a provider to define a mapping from one schema to another. It addresses such problems as mapping synonyms in the different schemas to each other, dealing with one to many and many to one relationships between the schemas, and providing a programmatic conversion from one value to another. The tool provides a graphic interface to define the mapping file, and other programs are used to automatically extract that metadata, apply the mapping that was generated to convert the data to the IIMS supported format, and then

ingest it into the IIMS clearinghouse. Through this process, new providers can participate in the clearinghouse. It is also planned to allow providers to participate in a distributed search, although those providers will lose the performance advantages of the clearinghouse.

The layered architecture of the IIMS provides a modular approach to providing a variety of interfaces into the system. As demand presents itself, simpler interfaces can be made available that will help drive new interface mechanisms. For instance, the existing query language could be used to perform a search based on geospatial and temporal parameters. However, to simplify the development of interfaces, a simpler query language that only allowed geospatial and temporal parameters with limited representation, could be provided. This is analogous to a novice versus an expert user interface provided on the programmatic interface.

Finally, it should be emphasized that IIMS interfaces are not all Application Programmer Interfaces (such as EJBs, CORBA or Java RMI). There are also interfaces that are more file-based. FTP can be used to place files in a directory that is automatically detected by the system and processed. The results are then made available in a similar fashion. This functionality will greatly simplify the creation of automated systems that interact with the IIMS.

CONCLUSION

The IIMS will make it possible for any data provider to easily participate in a science metadata clearinghouse so that their data can be shared with other scientists. Moreover, the IIMS will allow anyone to create their own user interface to access data held in the metadata clearinghouse. Special views of data as well as different data access paradigms will be supported.

REFERENCES

- [1] B. Schneiderman, "Dynamic Queries for Visual Information Seeking", Readings in Information Visualization: Using Vision to Think. S.K Card, J.D. Mackinlay and B. Schneiderman, Eds. San Francisco: Morgan Kaufmann, 1999, pp. 236-243.