

Next Generation Data and Information Systems for Earth Sciences Research

Liping Di¹ Ken McDonald²

¹NASA/RITSS

4400 Forbes Boulevard

Lanham, MD 20706, USA

E-mail: Liping.Di@gsfc.nasa.gov

²NASA Goddard Space Flight Center, code 423.0

Greenbelt, MD 20771, USA

E-mail: Kn.McDonald@gsfc.nasa.gov

ABSTRACT Advances in the ability to collect huge amounts of digital Earth data through remote sensing and the rapid expansion of applications of those data have placed greater requirements on data and information systems that manage and distribute the data to user communities. Recent advances in computer technologies make creation of a better system possible. In this paper, we discuss the shortcomings in current data and information systems, envision the next generation of such system, and suggest technologies for building such advanced systems. In our mind, the next generation system has to be distributed and intelligent, yet integrated. The system should greatly simplify locating and acquiring data and information and should fully automate the preprocessing and assembling, enabling data users to concentrate on analysis and applications rather than data preparation. Through such a system, users should be able to obtain data and information in ready-to-use form through their web browser, regardless where and how the requested data and information are archived. Even if the requested data and information can not be found in the archives, the system should be able to automatically generate and assemble the product for users from source data intelligently.

In order to build such a system, some fundamental problems have to be solved, including: 1) how to manage and access large, distributed, heterogeneous interdisciplinary Earth data and information resources over the Internet as an integrated, seamless intelligent system in real time; 2) how to extract domain-specific knowledge and information from the data in such a system intelligently and automatically, based on users requirements; and 3) how to provide users with object-based, content-sensitive spatial and temporal searches of and access to the data, information, and knowledge in the system. The advanced technologies necessary for making such system a reality include 1) object-based distributed processing; 2) interoperability standards and technologies for Geospatial data; 3) Data mining and automated information extraction; 4) machine learning and artificial intelligence, and 5) advances in World-wide Web, Internet infrastructure, and computer hardware. A proposed architecture for the next generation of data and information systems is discussed.

KEY WORDS Data and Information System, Open Distributed System, Artificial Intelligence, Earth Sciences

1. Introduction

What percentage of their time do scientists spend on locating and acquiring data and information, and then preprocessing and assembling them into analysis-ready form? It is probably large. For scientists working on multidisciplinary Earth System Science (ESS) research, the answer is probably more than 50%. In our case, we actually spend more than 80%. Suppose that you had an unlimited number of skillful and knowledgeable research assistants and associates who are doing all those preparations for you for free. How much increase in scientific productivity could be achieved; how many more new analyses and experiments could be conducted; and how much more new scientific knowledge can be discovered, even at the current level of resources? Suppose all those assistants and associates could be replaced by a distributed intelligent data and information system that would be ready to help you through your desktop computer's web browser any time you wanted? You would say

the system is great. This is exactly the next generation system should be.

2. Case Problems in Current Systems

Three significant features distinguish Earth System Science research from scientific endeavors in other scientific domains, 1) the research is multi-disciplinary; 2) the research needs the great amount of data and information and may be computational intensive; 3) the research regions may be micro (e.g., a leaf), field, local, region, continental, or global.

Normally, the processes of knowledge discovery in Earth System Sciences involve three consecutive steps in the data and information flow: 1) *Geoquery*, 2) *Geodata and information assembly*, and 3) *Geocomputation*. *Geoquery* is location and acquisition of data from the data repositories. The *geocomputation* is analysis and simulation of the complex Earth system using the data and information from the *geoquery*. *Geodata* and

information assembly assembles the data and information from data centers based on the needs of geocomputation.

Because of the multidisciplinary nature of Earth System Science, datasets from data centers are very diverse, and in many cases, the temporal and spatial coverages, resolution, origination, format, and map projections are incompatible. Scientists spend considerable time assembling the data and information into a form ready for analysis in the geocomputation step, even when the analysis is very simple. Although data and information systems are the core for supporting ESS research, current systems, which mostly facilitate only data search and ordering, fall far short of supporting multidisciplinary ESS studies. The next two subsections, describing some typical studies conducted by scientists authoring this paper, illustrate this point.

2.1. Test and Verification of a New Remote Sensing Algorithm

One of the key elements for studying Earth as an integrated system is to monitor the Earth's land, ocean, and atmosphere for a long period of time by satellite remote sensing. The development of algorithms for deriving the geophysical parameters from remotely sensed data is the key in remote sensing science. The major task of a remote sensing scientist is to develop algorithms. To test a new algorithm, one must assemble a test dataset. Because an increasing number of algorithms require multi-source data, which are typically archived at different data centers with incompatible spatial resolution, projections, and formats, it is estimated that more than 50% of the scientists' time is spent on assembling the test dataset. An example of the difficulties scientists have encountered in current information systems for a scientific endeavor is given below.

Recently, scientists have developed a new remote sensing algorithm in a funded study [Di, 1997; Chunhan, 1998]. The new algorithm uses both low spatial resolution passive microwave and the high spatial resolution optical remote sensing data for a global hydrological study. To test the algorithm, we selected the state of Nebraska as the test site because of the availability of ground truth data. We used AVHRR 1-km resolution data and SSM/I 25-km resolution data as the sources for the test dataset. This test dataset can be obtained by issuing this exact single query: "*Send me co-registered AVHRR and SSM/I data in HDF format for the whole state of Nebraska acquired in day time on any day with acquisition time less than 2 hours apart between SSM/I and AVHRR, cloud*

cover less than 20% in the AVHRR data, and vegetation not very dense". Similar queries are very common in Earth science research. However, to the best of our knowledge, no existing or planned information systems, other than the one envisioned in this paper, can properly satisfy such a natural language query and assemble the necessary data for us automatically.

It took less than one week for us to develop and debug the new algorithm. However, we actually spent more than two months finding data, ordering them, waiting for the arrival of the ordered data, and finally preprocessing, and assembling the data into the form acceptable for our program. As experts in satellite remote sensing, we knew the SSM/I 25-km Pathfinder global daily dataset is available at the National Snow and Ice Data Center at Boulder, Colorado, and the AVHRR LAC data covering the state of Nebraska are available at the EROS Data Center in Sioux Falls, South Dakota. Both centers provide a web interface to their information systems for browsing their catalog, obtaining text descriptions of the data, and placing an order. Each SSM/I data file contains one day's SSM/I data covering the globe with many gaps. Distributed in CD-ROMs, the data are in the cylindrical equal area map projection and in plain binary format. The AVHRR-LAC data are stored one orbit segment per file, in standard level 1b format and in sensor geometry. The catalog is searchable by specifying the geographic coordinates. Neither of these two sites supports on-line browse and subsetting of those two specific products. Because we want to find a pair of SSM/I and the AVHRR data that were acquired for the state of Nebraska fewer than two hours apart, and because the SSM/I data in any day have big data gaps between orbits, we have no way to determine which specific files to order. Therefore, we have to order all SSM/I global daily products from 1986 to 1991 and all AVHRR-LAC data that fully cover Nebraska in early spring or late fall for the same period. The result was that we received 28 CD-ROMs and tens of 8 mm tapes, which contain tens of gigabytes of data. But we really only need one pair of data that meet the query criteria listed above, less than 2 MB. In the next several weeks, we processed the AVHRR level 1b data into lat/lon coordinate systems using our AVHRR preprocessing software developed by a Raytheon scientist. Then the data were ingested into the ERDAS system, to be interactively browsed for finding the data pair we wanted. After that pair was identified, we had to subset the data file for the state of Nebraska, and reproject the SSM/I data into lat/lon so that both AVHRR and SSM/I could be co-registered. Since the algorithm only reads the data

in HDF format, we converted the data from ERDAS format to HDF format by a homemade software tool. Once the pair of co-registered data for the state of Nebraska was in HDF format, our hydrological program took less than 30 seconds in our SGI workstation to produce the result for the state of Nebraska.

2.2. Ecosystems' Reactions to Global Climate and Environmental Change

As we know, the formation and function of ecosystems are controlled by the environment. Therefore, environmental changes will cause ecosystem changes. Scientists developed a semi-empirical model for modeling the functional relationship between meteorological parameters and the vegetation status represented by the remotely sensed Normalized Difference Vegetation Index (NDVI) [Di et al., 1994]. The model can be used to derive some essential functions of an ecosystem, such as how effectively the vegetation in the ecosystem uses water, light, and heat. The model is a one-dimensional temporal model that can be run in low-end computers such as PCs.

In order to improve the scientific understanding of how terrestrial ecosystems respond to the climate and environmental changes and how their essential functions are affected, we have tried to find how the essential functions of major terrestrial ecosystems around the globe have changed in the past two decades by modeling the climate-ecosystem relationship with the model. For each ecosystem, the inputs to the model are the annual time-series meteorological daily records (precipitation and temperature), and annual daily NDVI data measured by the satellite under clear conditions for the location where the meteorological data are recorded. Therefore, for each year and each ecosystem there are 365 records of temperature and precipitation, and about 100 NDVI values recorded by satellite for the location where the climate record is taken, with the total amount of data to be less than 2 kb. If we consider evaluating how the approximately 40 major ecosystems mapped in the Olson global ecosystem dataset have changed in past 20 years, the total amount of data will be less than 1.6 Mb (20 *40*2kb). This amount can be easily handled by any PC.

However, preparing datasets for such as a study is not easy. Based on our knowledge, the Olson ecosystem datasets are available at the NOAA National Geophysical Data Center (NGDC), and the global weather station data are available at the NOAA National Climate Data Center (NCDC). The global daily NDVI for the past two decades can be provided only by the NOAA/ NASA Pathfinder land

dataset archived at the NASA Goddard Space Flight Center (GSFC). All three institutes have a web interface for browsing their catalogs and ordering data; however, there is no way currently to select the NDVI data on-line from the archive at GSFC using the weather-station location information stored at NCDC and the ecosystem datasets available at NGDC. Although we only need about 80 Kb (40 ecosystem classes*100 NDVI at clear condition/year*20 years *1 byte/NDVI) of NDVI data, we have to order the entire AVHRR global 8-km land daily product with a total size of more than 150 GB. Doing so represents 2 million folds of waste in data delivery and storage. With such huge datasets and both labor and knowledge-intensive data preprocessing and assembly, only a small group of elite scientists with large funding, expensive computers, and large numbers of research assistants can handle such volume of data.

3. Vision, Requirements, and Significance of the Next Generation System

Imagine an intelligent information system that can handle queries such as: *"For each Olson ecosystem class, send me, for the past two decades, the daily temperature and precipitation records measured at a representative location of the class and the daily mean of NDVI values measured by satellite under clear conditions for a 24*24 km² area surrounding the location"*, automatically locate and extract relevant data/information, preprocess and assemble the data, and deliver the data in analysis-ready form to the user. We believe the next generation of data and information systems should be intelligent enough to meet such a challenge.

It is our vision that all data/information finding, preprocessing, and assembling should be taken care of automatically by distributed intelligent information systems. In other words, step 1 and step 2 of geo-knowledge discovery should be fully automated. This will allow scientists to concentrate their energy on developing new algorithms and discovering new knowledge. This vision is also shared by many other scientists doing global change studies [Dozier, 1997]. With such distributed intelligent information systems, scientists will be able to enter requests similar to those illustrated in preceding section. These requests are entered in their web browser and in just seconds or minutes later, the requested data and information will be returned to the scientist's local computer ready for analysis.

In order to build such an ideal system, some fundamental issues have to be solved, including: 1) how to manage and access large, distributed,

heterogeneous interdisciplinary Earth data and information resources over the Internet as an integrated, seamless intelligent system in real time; 2) how to extract domain-specific knowledge and information from the data in such a system intelligently and automatically -- based on users requirements; and 3) how to provide users with object-based, content sensitive spatial and temporal search and accesses to the data, information, and knowledge in the system.

The next generation distributed intelligent information system envisioned in this paper will significantly enhance scientific productivity and may lead to deep and far-reaching scientific discoveries in Earth System Sciences. The system will make it much easier and faster for scientists to investigate a wide range of earth related problems from health to soil moisture, which the present earth data archive may be able to answer. It will also make it possible for fully exploring and making effective use of the huge amount of Earth science data to be collected by US federal government's global change programs in the coming years. In addition, the system also makes possible many studies from scientists' desktop computers that currently are impossible. This will transfer the multidisciplinary Earth System Science Studies from a small elite group with plenty of resources to many more scientists equipped with only cheap web-capable PCs.

4. New Technologies Available for Such Ideal Systems

In recent years, several key technologies have been advanced significantly and are mature enough for use to build the next generation system envisioned in this paper. The technologies include: 1) object-based distributed processing; 2) data, metadata, and spatial/temporal object interoperability through FGDC, ANSI, ISO, and OGC standards; 3) Data mining technologies; 4) machine learning and artificial intelligence; and 5) advances in World Wide Web, internet infrastructure, and computer hardware. By integrating those technologies, the next generation system can be developed.

4.1. Object-based Distributed Processing

Technological advancement has made the object-based distributed processing possible now. The technologies include the object-oriented platform independent code system (e.g., Java) and distributed object infrastructure (e.g., CORBA, OLE/COM). With those technologies, a truly distributed system with interoperable processing can be constructed.

4.2. Geospatial Data Interoperability Standards and Technologies

In order to make the geo-processing interoperable in the distributed environment, the object-based distributed processing technologies are not enough. Domain specific objects and protocols have to be defined and standardized. In recent years, national and international standard bodies have developed geospatial data, metadata, catalog, and function standards. The major standard bodies actively developing geospatial data and interoperability standards include the Federal Geographic Data Committee (FGDC), the American National Standard Institute (ANSI), the ISO TC/211, and the Open GIS Consortium (OGC). Implementing those interoperability standards with object-based distributed processing technologies will form the framework for the next generation data and information systems. On this framework, intelligent functions can be developed for automated geospatial data finding, assembling, preprocessing, and knowledge discovery in the distributed environment.

4.3. Data Mining Technologies

Data mining deals with the discovery of hidden knowledge, unexpected patterns, and new rules from large databases [Glymour et al. 1996, 1997, Glymour, 1995, Gray et al. 1997]. It is one of the key components in the process of information and knowledge discovery in databases. Examples of the technologies associated with data mining include the content-based search, statistical techniques, data and information visualization, decision trees, association trees, neural networks, and genetic algorithms. Data mining technologies have been studied intensively in recent years for applications in the digital libraries and data warehouse. Prototypes of such technologies have been successfully demonstrated. It is the right time for using those technologies in the next generation data and information system for Earth Science research.

4.4. Machine Learning and Artificial Intelligence

With the exponential growth of Earth science data from satellite remote sensing, it would be most useful to use intelligent methods for automatically extracting knowledge on a timely basis and to decide where and to whom selected subsets of data should be distributed. Among some current intelligent methods of increasing promise are (1) artificial neural networks (memory-based and layered architectures) [Surkan and Di, 1989; Straler et al., 1996; Surkan and Campbell, 1997], (2)

genetic algorithms (standard, fixed-length and messy, variable-length chromosomes) and (3) approximate reasoning methods such as fuzzy logic rule-based systems.

It is becoming evident that intelligent systems will eventually be designed to capture the expertise of human data managers and analysts. This will eliminate human intervention between producers and users of data. To this end the current human support links will become more automatic and transparent. In the history telephone systems and the current use of earth systems data, there is an analogy with the situation in which the role of the human switchboard operator vanished because new call-routing software. This software has become so knowledge rich that it is now practical to require that a user obtain the service or contact the person desired directly without any intervention by a human expert.

It is expected that intelligent systems must be developed for ESS to facilitate the users in accessing, by themselves, the information or data they require. Such knowledge-rich systems must also collect statistics for the current and projected demands for services. Such statistics can be used for automating the management of the data acquisition and delivery systems. Constructing such systems will require the discovery and defining of the rules for data management. Also, we must encapsulate the operators or functions that transform data from one format or representation to another. Intelligent agents must communicate through dialogs with users and devices. These dialogs will maintain contact and perform required functions at acceptable times.

4.5. Advances in World-wide Web, Network, and Computer Hardware

The next generation system envisioned in this paper has to be web-based because of popularity and efficiency of this technology. However, web technology is changing very rapidly. A number of new standards proposed by W3 consortium would have a significant impact on the development of data systems in the near future. The Extensible Markup Language (XML) will be the next generation language for web. XML helps to understand the data content standard and will complement HTML. Extensible Style Sheet (XSL) is another standard that works with XML helps to render data in multiple ways. There will be discipline specific standards in addition to XML and XSL such as ChemML for chemistry, Math ML for Mathematical representation of data. Resource Description Framework (RDF) will help in the organization and

description of metadata. RDF will help in improving data system design and search.

These new web standards will improve the search results significantly. Manipulation and display of results in multiple ways on the web will become a reality. Intelligent search for information and complex queries can be implemented in a data system. Building links between various types of information will become possible. Data interoperability will significantly improve through Document Type Description (DTDs).

The next generation Internet provides much faster data communication than current Internet. The next generation Internet will be the key network infrastructure for the next generation data and information system because the need for communicating the large volume of Earth science data over the network.

The data mining and artificial intelligence algorithms typically required significant amount of computational power. The rapid advances in the computer hardware make the use of those algorithms in the operational environment feasible. In addition, running programs on new, faster computers now make practical, the application of computation intensive resampling statistics of the types developed by Efron at Stanford beginning in the 1980's. The use of these newer statistical methods can now exploit computers to generate high-resolution pictorial displays (colored graphics) which make their meaning more obvious. Until the appearance of the present computers and very high resolution displays, this has not been so straightforward. The time is now ripe for many potential useful display formats for data and results to be conceived and tailored to specific classes of data. Graphic displays can now be generated with a resolution that exceeds the limits of human visual acuity. It now possible to create displays to use both disjoint and overlapping color maps. Such displays can be tailored to highlight data trends and systematic internal relationships. The objective is to make such internal structure easier to detect by visual inspection and to complement standard data analysis procedures. Hybrid interactive programs can exploit the full processing power that is possible from the combined application of the human brain and visual system. It is desirable to make it easier to discover predictive data structure even in the presence of occluding or superposed noise.

5. A Proposed Architecture for the Next Generation System

We consider a granule of geoinformation (either a dataset, a query result, or geocomputation output which describes some aspects of Earth) to be a

geo-object, which consists of data itself, a set of attributes (metadata), and a set of methods that can be operated on the object. Of course, one of the methods is the object creation method itself, which may consist of methods from the parent objects. A geo-object stored at a data center is an *archived geo-object*. Therefore, raw datasets sensed by remote sensors or the field investigations would be the root geo-objects because no other geo-objects can be used to create those objects. All geoinformation and knowledge are child objects of those root objects. Thus, from object point of view, all processes for geo-information/knowledge discovery are the processes of creating new geo-objects. If we consider a user request is a user-defined geo-object, or called user geo-object, in most cases, there is no archived geo-object available that exactly matches the user geo-object. *The task of the next generation system is to automatically construct the user geo-object from the archived geo-objects to fulfill users' request in a distributed environment.* Intelligence is associated with the processes of finding the roadmap from the user object to archived geo-objects, and constructing the user geo-object automatically from the archived geo-objects. The knowledge discovered through this process is represented in the user geo-object.

In the following subsections we will describe a proposed next generation system that will automatically generate the user geo-objects at users' request. This will be done with intelligent and automated methods in the distributed environment. In order to simplify the description, we subsequently call the proposed next generation system the *GeoBrain*.

5.1. Distributed System Architecture

The next generation system will be Web client-server based intelligent data and information system capable of both working alone and forming a federation of providers that use GeoBrain and other information systems with standard interoperability protocols. On the server side, GeoBrain will consist of 1) an intelligent geoquery interpreter to convert a geoquery sent by users (a user geo-object) into a set of geo-objects (first-level object decomposition, we call that level of objects the immediate objects) with required attributes and the relationship among the objects; 2) an intelligent information broker that will work with an intelligent geo-search engine to form the roadmap from the user request geo-object to the archived geo-objects; 3) an intelligent search engine that will quickly and effectively locate the requested object in the vast archive; 4) an intelligent geo-object

assembler that will produce the user geo-object from the archived geo-objects based on the roadmap created by components 2 and 3, regardless where the archived geo-objects are located; and 5) a presentation manager that will convert the user geo-object into the form required by the user. In the client side, the interface to GeoBrain will be any Java capable web browser. A graphic user interface will be provided to users for interactively inputting a request and manipulating the search results. An optional application program interface may be built so that the query can be initiated by an analysis program and the results returned to the program for direct analysis. The individual components will be explained in following sections.

5.2. Geoquery Interpreter

The primary task of the geoquery interpreter is to understand what the user is requesting and to convert the request into the form that GeoBrain can understand, a formal geo-object description. It is true that we can consider that every request from a user can be met by a geo-object. However, we have to convert the user's request to the formal geo-object description accurately.

There are two types of user geo-objects described by a user query, simple and composite. A user object is simple when only one parent object is requested explicitly in the query, such as "request annual maximum NDVI for USA in 1996", where annual maximum NDVI is the parent object of the user object. A composite user geo-object contains more than one parent geo-object, such as that described by the query in section 2.1. Regardless of the kind of user geo-object, the geoquery interpreter has to convert the user request description into (1) a set of immediate geo-objects that are the parents of the user geo-object, (2) a set of required attributes, and (3) the relationships between the user geo-object and the immediate geo-objects.

In order to make entering of requests flexible for users, a natural language query will be the primary form for transmitting users' requests to GeoBrain. Of course, certain limitations will be imposed on the grammar of query structure. However, the interpreter has to be intelligent enough to understand the semantics of the user's query and the explicit and implicit conditions carried within the user's query. Therefore, the interpreter must have the ability of natural language understanding and the domain-specific knowledge about the geo-object. For example, for the request given in section 2.1, the interpreter has to convert the request to two immediate geo-objects: an SSM/I object and an AVHRR object. The required attributes for both geo-

objects are data format, acquisition time, map projection, and spatial coverage. The additional required attributes for AVHRR are percentage cloud cover and percentage vegetation cover. The method for creating the user geo-object from the two immediate objects is a selection operation that selects the pair of datasets satisfying the condition. Therefore, the interpreter must understand that the selection operation is implied although it is not explicitly stated in the request. In addition, the request also has qualitative condition such as the vegetation being not very dense. Domain-specific knowledge has to be presented in order to understand these qualitative conditions. In here, the condition can be met by either finding winter, early spring, or later fall's images or converting this qualitative condition to a quantitative one such as the mean NDVI value is less than 0.1.

5.3. Intelligent Information Broker

The task of the information broker is to work with the intelligent search engine to find the roadmap between the archived geo-objects and the user geo-object in following ways. (1) A request is sent in by the query interpreter locally or from a remote site in a federated and distributed environment. Note that a query from the query interpreter may consist of multiple immediate geo-objects while that from a remote site consists of only one geo-object that is one of the ancestors of the user geo-object. (2) The broker will find which information providers are most likely to have the requested geo-objects in storage and send the request onward to probable providers. (3) If one of the requested objects can not be found, it will be decomposed to its parent objects and repeat step 2 again. Step 2 and Step 3 form a loop that will be terminated only when the requested object is constructed or the object can not be constructed from the existing data and information in the federation of GeoBrains. All decomposition of geo-objects and their descriptions will be kept for constructing the user object later. This information can be kept in an inverse tree structure, which we call *geotree*. The root nodes of geotree are geo-objects archived at distributed GeoBrain sites and the leaf is the user geo-object. The tree will be used as the roadmap to construct the user geo-object from archived geo-objects in the geo-object assembler. The query sent in from a remote site will result a sub-geotree, which will be sent back to the remote site to form the whole geotree.

To fulfill the above tasks, the information broker needs four sub-components: a geo-object pathfinder, a geo-object decomposer, a geo-object tracer, and an interoperability manager. The geo-object pathfinder has the knowledge of which sites

in a federation of data providers has highest probability to find the needed geo-object. In order for the pathfinder to know where the required geo-objects for generating the user geo-object are located, high level catalogs (such as the collection level) must be exchanged between the federated sites by using protocols and metadata standards such as those of ISO or FGDC through the interoperability manager. The geo-object decomposer will decompose a geo-object into its immediate parents based on the domain specific knowledge about what kind of input objects can be used to generate the geo-object and how it is generated (e.g., for an NDVI object, we know that we need a red reflectance object, a near-infrared object, and an NDVI creation algorithm). Both pathfinder and decomposer will be implemented by using a fuzzy logic expert system with a domain-specific knowledge base. The geo-object tracer will record all steps of geo-object decomposition and form the geotree.

The interoperability manager will manage all communication with other data providers' sites in a distributed environment. There are several interoperability protocols in the geo-information area. The ANSI Z39.50 standard for information retrieval is a generic application program interface (API) to interact with any database or information service [ANSI/ISO, 1995]. The Committee on Earth Observation Satellites (CEOS) Working Group on Information Systems and Services Protocol Task Team (WGISS/PTT) defined a Z39.50 profile called the Catalogue Interoperability Protocol (CIP) for locating Earth Observation data from any data provider [CEOS, 1996]. The Federal Geographic Data Committee (FGDC) has also developed a Z39.50 profile, called the Geo Profile, for interoperability among FGDC Clearinghouse nodes. We propose to implement the Z39.50 in our GeoBrain system using Java. In addition to CIP and Geo Profile, NASA EOSDIS also defines several interoperability protocols, such as EOSDIS EDG and ECS protocols. The Open GIS Consortium (OGC) has also defined interoperability protocols, such as Web Mapping Specifications. The interoperability manager of the GeoBrain should understand the commonly used protocols.

Users can discover the contents of a particular repository by querying its catalog (inventory) service. With relevant scientific data located at hundreds or thousands of repositories around the world, it is crucial to be able to locate data wherever it may exist by just querying one site--a one-stop shopping approach. The intelligent information broker will provide GeoBrain the federation and one-stop shopping capacities. A GeoBrain site can

interoperate with other GeoBrain sites or with sites with the same protocols as GeoBrain to form a single, federated information space.

5.4. Flexible Search Engine

The intelligent search engine will accept the search request from the information broker and then perform a two step search. First, it will search the catalog database to find if there is an archived geo-object that possibly meet the request (Note: the request consists of a subset of the user object). Second, it will do the content-based search to find if the archived object exactly matches the request.

The two step search is necessary because of the characteristics of the Earth system science information. First, the archive is normally huge and geo-objects in the archive very diverse. The catalog search will eliminate a lot of irrelevant geo-objects from slow content-based search, improving the overall performance of the system. Second, the archived geo-objects almost always have larger spatial and temporal coverage than the requested geo-object. Because the geographic region requested by a user is arbitrary (i.e., GeoBrain does not know in advance which regions in the globe the users want), the metadata in the catalog will be global in nature describing features pertaining to the whole coverage, and the pre-indexed contents are relevant only to the whole coverage. Therefore, the specific content pertaining to a region has to be searched by content-based search. For example, suppose a GeoBrain site has a set of daily NDVI datasets and a set of monthly NDVI datasets, both covering USA from 1981 to 1993, and there is a request for daily NDVI datasets for Nebraska for June 1986 with less than 10% cloud cover. The first step search will eliminate those daily NDVI datasets that are not within the month of June 1986 and all monthly NDVI datasets. Although the metadata for the archived NDVI datasets may contain the cloud cover information, it describes the whole archived geo-object. Cloud may cover 50% of the archived geo-object but only 5% of Nebraska. Therefore, 10% cloud cover can not be used as a first-step search criterion, while the spatial and temporal criteria have to be used. The search engine must be able to recognize which search criteria belong to the first-step search and which to the second step.

The GeoBrain should have the capabilities of both the catalog-based and content-based searches. The catalog-based search is very common in the current generation systems, such as Data and Information Access Link (DIAL) [Di, et al., 1997, 1999]. Content based search methods will need to be developed for GeoBrain. There are two alternative approaches, physical and non-physical, in content-based search.

In GeoBrain, each content search always has physical meaning and can be derived from physical-based algorithms. For example, a lot of geo-objects have a cloud mask for each pixel. For such case, we can directly count the pixels within a region to get the region-specific content information. If such information is not available for a geo-object, we always can derive it from the parent geo-object by using a physically based algorithm. For the cloud case, we can use the temperature thresholding algorithm in a thermal infrared band to get the cloud mask. However, in many cases, the geo-object on which the content-based search is performed is a root geo-object, therefore, the required content may not be derivable. For non-physical methods, we are looking for patterns in a geo-object. Statistics-based, neural network, and pattern recognition methods are examples of the non-physical based methods.

5.5. Smart Geo-Object Assembler

After the roadmap from the user geo-object to the archived geo-objects is formed, the geo-object assembler will automatically assemble the user geo-object from the archived geo-objects by using the roadmap (the geotree) and the geo-object creation methods, and archived geo-objects. Assembling the user geo-object is a large task that may require significant computation power. To assemble a user geo-object, many intermediate geo-objects may need to be created, depending on the complexity of the user geo-object. Very commonly, a geo-object is created through a combination of multiple parent geo-objects by a geo-object creation algorithm. In a distributed environment, those parent objects may not reside in one GeoBrain site. Therefore, a key issue has to be solved for the next generation system, the dynamic task allocation.

The purpose of dynamic task allocation is to reduce the overall time and resources needed for assembling the user object by assigning part of the assembling work to sites that have geo-objects to be used and resources available for such work. Two criteria have to be used in such an assignment; minimizing the overall assembly time and minimizing the network traffic. As an example, for the request in section 2.2, we don't want to assemble the user geo-object at the site where the Olson ecosystem object resided even if the user's query was initiated from the site because shipping the NDVI datasets through the network is too time consuming. In such a case, the assembling work will be performed by the site where the NDVI resided.

One issue in dynamic task allocation is method interoperability. If a geo-object assembling task is assigned to a site, we have to move the object creation method to the site also if the method is not already available there. In the distributed environment, this method may not run because of the difference in operating system and hardware. Therefore, platform independent languages, such as Java, have to be used in encoding the geo-object creation methods. Another issue is how to ship the geo-object between sites without losing information. This will require the interoperability at the geo-object level. Several candidate technologies, such as the Open GIS Consortium's simple feature protocol, can be implemented in Java, ActiveX, and CORBA.

5.6. Presentation Manager

After the user geo-object is assembled, the presentation manager will present the geo-object to the user through the graphic user interface or the optional application program interface (API). The presentation manager will enable users to manipulate the user geo-object interactively and flexibly. Part of this component has already been implemented in many current systems, which can be reused in GeoBrain. Once issue needed to be solved will be how to present the results of a complex query to user intuitively and informatively. Multi-media briefing may be a way to pursue, presenting the user geo-object in graphics, images, text summary, and possibly voice.

5.7. Interactive User Interface

GeoBrain should have two user interfaces. The main one is the Web-based graphic user interface through Java capable web browser. In addition, an application program interface should be provided so that an application program at user's computer may directly interact with GeoBrain. The graphic user interface will permit data users to enter queries and view and manipulate results interactively with the presentation manager. The Java based interface will run in Web browsers on the client side as applets, providing fast interaction and prompt response.

5.8. Candidate Intelligent Methods for GeoBrain

In GeoBrain, we need to develop many geo-object creation and processing methods (algorithms). There are two types of methods, *global* and *domain specific*. Global methods, such as reprojecting, co-registering, resampling, reformatting, and subsetting, can be used in many geo-objects and have mature algorithms. The domain specific

methods, such as deriving a leaf area index from NDVI, can only be applied to domain specific geo-objects. There are many domain-specific methods available currently at NASA. These methods can be integrated into GeoBrain.

In addition, there are some novel intelligent methods important to satisfying the selection and processing of Earth science data. These methods may be implemented in the GeoBrain system, depending on their relevance to the creation and manipulation of individual geo-objects. Among some current methods of increasing promise are (1) artificial neural networks, (2) genetic algorithms, (3) approximate reasoning methods such as fuzzy logic rule-based systems, and (4) non-parametric techniques such as the computing-intensive methods that originated with the bootstrap and jackknife statistical techniques.

6. Time Frame for the Development of the Next Generation Systems

The above mentioned technologies are mature enough for being integrated and implemented in the next generation data and information systems for serving the Earth science research. It is expected a prototype of an intelligent distributed information system, such as the one proposed in this paper, can be available within three years. Such a prototype system will demonstrate the power of the integration of those new technologies in providing data, information, and knowledge services to Earth science research. Such a prototype system can be further developed into an operational data and information system quickly. We expect that the next generation system will be in operational uses within six years from now.

References

- ANSI/NISO, 1995, Information Retrieval: Application Service Definitions and Protocol specification: Z39.50.
- CEOS, 1996, Catalogue Interoperability Protocol (CIP) Specification - Release B, [ftp://harp.gsfc.nasa.gov/incoming/ptt/cip-b201.ps]
- Chunhan, 1998, Soil Moisture Algorithm Theoretical Basis Document for the Visible and Infrared Image Radiometer Suite (VIIRS) onboard the National Polar-orbiting Operational Environmental Satellite System (NPOESS), Raytheon ITSS Research Report submitted to Intergovernmental Program Office for NPOESS, Silver Spring, MD.
- Di, L., R. Suresh, K. Doan, D. Ilg, and K. McDonald 1999. DIAL-an Interoperable Web-based Scientific Data Server, In M. Goodchild, M. Egenhofer, R. Fegeas, and C. Kottman edit, *Interoperating Geographic Information Systems*, Section 4. System Experiences. Kluwer Academic Publishers, Norwell, MA, pp 301-312.

- Di, L., 1997, Derivation of High Spatial Resolution Soil Moisture Information from the Combination of Optical and Microwave Remote Sensing Data, Raytheon ITSS Research Report, Lanham, MD.
- Di, L., R. Suresh, K. Doan, D. Ilg, and K. McDonald, 1997, A Web-based Scientific Data Server for Accessing and Distributing Earth Science Data, *Proceedings, International Conference on Interoperating Geographic Information System (Interop' 97)*. December 3-4, Santa Barbara, California USA. pp 243-254.
- Di, L., Rundquist, D. and Han, L., 1994, Modeling Relationships Between NDVI and Precipitation During Vegetative Growth Cycles, *International Journal of Remote Sensing*. Vol. 15, No. 10 2121-2136.
- Dozier, J., 1997, Information system for global Change Study, Keynote speech at the 1997 International Conference for Interoperating Geographic Information System. Santa Barbara, CA.
- Glymour, G., M. Madigan, D. Pregibon, P. Smyth, 1997, Statistical Themes and Lessons for Data Mining, *Data Mining and Knowledge Discovery* 1(1): 11-28.
- Glymour, G., D. Madigan, D. Pregibon, P. Smyth, 1996, Statistical Inference and Data Mining, *CACM* 39(11): 35-41.
- Glymour, C., 1995, Available Technology for Discovering Causal Models, Building Bayes Nets, and Selecting Predictors: The TETRAD II Program, *KDD* 1995: 130-135
- Gray, J., S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, H. Pirahesh, 1997, Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross-Tab, and Sub Totals, *Data Mining and Knowledge Discovery* 1(1): 29-53.
- Strahler A., J. Townshend, D. Muchoney, J. Borak, M. Friedl, S. Gopal, A. Hyman, A. Moody, and E. Lambin , 1996, MODIS Land Cover Product Algorithm Theoretical Basis Document (ATBD) Version 4.1. <http://tpwww.gsfc.nasa.gov/MODIS/MODIS.html>
- Surkan, A.J., Colin Campbell, 1997. - WWW publication and CD-ROM Archive of paper titled: "Constructive Algorithm for Neural Networks that Generalize" SIGAPL APL97 Proceedings of the Conference with Theme: "Share Knowledge, Share Success" held at Toronto, Canada August 17-20 {refer to the cd-rom, root directory for INDEX.HTM and TOOLS.HTM files for pointers}
- Surkan, A. and Di, L., 1989, Fast Trainable Pattern Classification by A Modification of Kanerva's SDM Model, *Proceedings of First International Joint Conference on Neural Networks*, International Neural Network Society, Washington D. C., pp. I-347-I-349.