

**The University of Southwestern Louisiana
Energy and Environmental Technology Applications Program
Information System Technology Project**

A Progress Report

DOE Grant No. DE-FG02-97ER12220

Project Officer: Dr. Samuel Barish

Task Monitor: Dr. Dennis Traylor

Project Managers:

Gaye Farris

National Wetlands Research Center

U.S. Geological Survey

Vijay Raghavan

Center for Advanced Computer Studies

The University of Southwestern Louisiana

24 November 1998

1 Overview

The objectives of this project are: (i) to establish an Energy and Environmental Information Resources (EE-IR) Center; (ii) to conduct research work in Internet computing in support of the EE-IR Center; and (iii) to enhance the facilities of the NASA/USL Regional Application Center (RAC). The project is carried out at the University of Southwestern Louisiana (USL) by faculty, staff, and students in the Center for Advanced Computer Studies (CACS). Through this work, we have collaborated with DOE's Office of Scientific and Technical Information (OSTI), and with the U.S. Geological Survey's National Wetlands Research Center.

The EE-IR Center is developed in collaboration with the National Wetlands Research Center (NWRC), and with OSTI. In December 1997 and January 1998, we had two meetings, and identified and interacted with potential end-users of the EE-IR Center and the GIS community, respectively. While we are working with NWRC on data collection and ingest issues, we are also developing software tools addressing different aspects of digital library development. We have studied the strategies for enhancing pre-existing Web documents in order that just the relevant portions (or components) of a document can be retrieved. We have also developed an approach for searching product catalogs on the Internet, based on concepts, so that an intermediary can conduct the search on different databases. Our approach to concept-based retrieval should be of interest to refining certain kinds of canned searches that are often made in the DOE Information Bridge.

Developmental work for the EE-IR Center can be classified into two tracks: work by the library group on metadata, cataloging process, and workflow; work by the software tools development group. The library group consists of Judy Buys (NWRC reference librarian), Adam Chandler (systems librarian), Dan Foley (metadata librarian), and Suzanne Harrison (GIS specialist). They have been working since April 1998 on data collection for the EE-IR Center. Software tools development is performed by Vijay Raghavan (CACS faculty; project co-manager), Alaaeldin Hafez (research scientist), Marie Erie (research scientist), Srikanth Koritala (research scientist), and a team of CACS graduate students. The development of tools is performed in conjunction with the library group, and represents an integrated team effort. These two tracks are described in more detail in the following.

Research work in Internet computing is conducted primarily at CACS by Vijay Raghavan, Henry Chu (CACS faculty), Kemal Efe (CACS faculty), Alaaeldin Hafez, and a team of CACS graduate students. Our research work is conducted in support of the development of EE-IR Center in the following areas: multimedia retrieval; data mining; parallel and distributed computing; media technology and data visualization.

The RAC provides a source of imagery, text, and geospatial data for the EE-IR Center. Participants from the RAC are Brent Yantis (RAC director) and Nathan Handley (RAC projects manager). Another source of image and geospatial data is Larry Handley (NWRC geographer) and his colleagues from the NWRC spatial analysis branch.

2 Library Development

The library group's work is to locate data in numeric, text, and graphical display and to

document the data in a standard Federal Geographic Data Committee (FGDC) or National Biological Information Infrastructure (NBII) format. Within the subject areas of energy and the environment, there is a special focus on data about pollution and contamination. Geographically, the emphasis is on Coastal Louisiana, the Lower Mississippi Valley, and the Gulf of Mexico. The metadata will be served on the Internet in standard FGDC or NBII formats. In addition, the metadata will be mapped to Machine-Readable Cataloging (MARC) for inclusion in the data server at the center and an international data server, WorldCat. Data served by the EE-IR Center will be added to a library of data using collection development decisions compatible with the scope of the EE-IR Center. Data will initially be documented with Dublin Core elements and then those elements will be mapped to MARC fields and NBII metadata standards.

2.1 Metadata Issues

Metadata, or structured "data about data," is used to discover, describe, and index the digital library collection. Three kinds of metadata are used: (i) Dublin Core metadata for Web site data; (ii) FGDC/NBII metadata for geospatial and biological data; and (iii) USMARC.

MARC, which stands for Machine Readable Cataloging, refers to a set of related formats for bibliographic description developed by the Library of Congress in the 1960's. These computer readable formats are based on the ISO standard 2709. They form the basis for all library cataloging done in the United States. MARC records are entered into WorldCat, a bibliographic database maintained by OCLC, the Online Computer Library Center, in Dublin, Ohio. WorldCat currently has over 40 million USMARC records created by 27,000 member libraries worldwide. A subset of WorldCat called InterCat contains USMARC records for Internet resources.

Dublin Core metadata is being used in this project because it is specifically designed for the description and location of Internet resources. In fact, Dublin Core is emerging as the international standard for Web site cataloging and is currently being used in over twenty countries besides the United States. The Dublin Core has fifteen elements as shown in Table 1. Each has a clearly defined meaning as a result of consensus reached at international meetings. Nevertheless, Dublin Core is also evolving to meet the changing needs of the Internet community. For example, the entry "Coverage" refers to spatial and temporal data. In addition to place names and dates, a working group is revising this element to include entry of latitudes and longitudes. Currently, there are just over 100 Dublin Core metadata records at the EE-IR Center Web site (the URL at <http://eeirc.nwrc.gov>).

The Dublin Core record is central to our cataloging process. Once an item has been selected to be included in the EE-IR Center's collection, it is searched in WorldCat to determine if it has been catalogued by another library. If so, this record can be edited to create a Dublin Core (DC) record. Most of the time (over 90% of cases), however, original DC metadata is created

Table 1 The fifteen elements of the Dublin Core.

Title	Contributor	Source	Creator	Date
Language	Subject	Type	Relation	Description
Format	Coverage	Publisher	Identifier	Rights

and entered in a template. This creates a DC record in text and Hypertext Markup Language (HTML) formats. The HTML record is attached as a header to the Web page in cases where a pre-existing page is also being added to the local library collection. At present, a DC template at the Nordic Metadata Project site, located in Sweden, is being used, but a DC template for the EE-IR Center is being developed. Part of the cataloging process will be to create a USMARC record from the DC record and to upload it into OCLC's WorldCat. Like the DC template, this is still under development, and it requires a DC/USMARC converter. There is a paper crosswalk between Dublin Core and USMARC written by the Library of Congress. Using this crosswalk, EE-IR Center librarians are working with the software development team to develop the template and converter. Once this is done, Library metadata records will be available internationally via both the Internet and WorldCat.

The scope of the collection of spatial data is primarily limited to coastal Louisiana. Secondary data needs are for the entire state of Louisiana and the Gulf of Mexico coastal states. Surveys for data and metadata are coordinated with the Louisiana Geographic Information Center (LAGIC), the NWRC, and the RAC. The EE-IR Center is documenting data and information sources using national standards outlined in the Content Standard for Digital and Geospatial Metadata (CSDGM) accepted by the Federal Geographic Data Committee (FGDC). This standard enables the sharing of spatial data among producers and users. Metadata can be documented using 334 potential elements in the broad categories shown in Table 2. The objectives of the standard are to provide a common set of terminology and definitions for the documentation of geospatial data and to provide a standard guide to the potential user to determine if a data set is appropriate for the user's intended purpose. The standard provides a way for data users to know what data are available; whether the data meet specific needs; where to find the data; and how to access the data. This will help avoid wasteful duplication of efforts and ensure effective and economical management of information resources in meeting essential user requirements.

The CSDGM clearly defines how a data set containing geographical referenced data will be documented. As outlined in the Executive Order 12906, federal agencies are required to use the standard and make sure that the data they are collecting do not already exist and are available elsewhere. While the standard defines what elements are used, it does not define how the data will be served or what software will be used to display and search the data. Many sites use the ISITE software, which is freely available on the Internet from the Center for Network Information and Discovery and Retrieval (CNIDR) and which is Z39.50 compatible. This allows each agency to manage its own data, and the Z39.50 standard allows a common interface. The EE-IR Center will use ISITE to serve data as a node on the NSDI.

Besides geospatial data, there are many data sets and information sources that are non-geospatial. These include much of the biological data such as habitat inventories, laboratory

Table 2 Categories of the 334 potential elements in the FGDC Content Standard for Digital and Geospatial Metadata.

Identification Information	Data Quality Information
Spatial Data Organization Information	Spatial Reference Information
Entity and Attribute Information	Distribution Information
Metadata Reference Information	

observations, biochemical characterizations, physiological observations, and bibliographical data. Biological data can have a geospatial character and these include vegetation maps, museum records for specimens, and monitoring data for animals. To make the current FDGC standard flexible enough to use for biological data, the USGS Biological Resources Division (BRD) proposed the Metadata Content Standard for Biological Resources Data. BRD is charged with fostering a cooperative effort to share biological information, and they are doing this through a confederation of data bases offered through the National Biological Information Infrastructure (NBII). The NBII is a node on the National Spatial Data Infrastructure (NSDI) and also uses the ISITE software. The standard does not have an Executive Order mandating its use. The biological standard is a special version of the FGDC standard and has additional fields for taxonomic references, a description of the analytical tools needed to interpret the data, and a description of the field or laboratory methodologies used to collect data. Therefore the biological standard may contain both geospatial and non-geospatial data. The EE-IR Center is using this biological standard to document biological data sets.

To make data entry even more standardized, there are FGDC standards committees trying to standardize terminology. The FGDC has approved use of the Spatial Data Transfer Standard, the Cadastral Data Content Standard, and the Classification of Wetlands and Deep Water Habitats of the U.S. for use with the accepted version of the CSDGM. Additional standards for vegetation classification, soils, biologic nomenclature and taxonomy are still in review. The EE-IR Center is following the progress of these standards and will implement them as they are approved.

2.2 Tools and Workflow

2.2.1 Collection Development Decision Process

There are two broad categories of resources collected and described with metadata by the EE-IR Center. One category is materials which are already on the Web. For these, we use the 15-element Dublin Core metadata standard. The second category is raw data which at the federal level is mandated that the FGDC metadata standard be used to describe it.

Every library needs a policy for deciding what content will be included in the collection, and what will be excluded. Also, as in any library, decisions must be made as to what materials to archive locally, and which materials will be linked to or retrieved from a remote site. As a general rule, we are downloading a local copy of documents whenever there is a chance that the link might not be maintained in the future, and if the document will not be changed in the future. So, for example, we point to some weekly serials produced on the energy industry by the Department of Energy, because it is more efficient, and more accurate, than having the EE-IR Center staff having to go to the site each week and downloading the file. On the other hand, the DOE has produced some data tables on the energy industry which are arranged by state, for the period of 1970-1995. These kind of historical data are not dynamic—unlike a weekly serial. Therefore we made the decision, especially since the files are small, to point to a local copy in such cases to insure availability. By writing metadata for the file, and pointing directly to it, we add value for the user who now can go directly to the information, rather than wading through a foreign Web site. We are making these historical energy files available in two formats: text

(for importing into a spreadsheet), and pdf (for easy printing and viewing). For example, see the URL at <http://eeirc.nwrc.gov/metadata/19.htm>.

Another part of this decision process involves selecting resources from within a Web site. We have basically three levels for Web resources. The first level is Dublin Core metadata that describes the collection of Web pages of an organization. The second level is metadata for a section or page within the Web site. The third level is for a specific database, document, or data table. We make this decision based on the quality of the materials, and its relationship to our collection policy. For any particular source, we may have all three levels, two levels, or one level. The librarians and our GIS specialist work through this process for sites brought to our weekly meeting.

Our FGDC metadata effort is by necessity slower. The above mentioned FGDC standard is very complex, containing over 300 elements. Our GIS specialist is working with this standard. One of our considerations is to be aware, in some detail, of what is happening at the state level in Louisiana, in particular, with the LAGIC group at LSU and Louisiana GIS Council (<http://atlas.lsu.edu/lagic/>) so that we do not duplicate the state agencies' efforts. Therefore, we are working our way outward from resources at USL and NWRC, then on to federal and other local data sources that fit our collection development policy.

2.2.2 OCLC and InterCAT

InterCAT (<http://orc.rsch.oclc.org:6990/>), as of 19 November 1998, contains some 57,000 items. We check InterCAT before we write metadata to prevent duplications. Our records will be added to InterCAT in the near future.

2.2.3 Dublin Core Metadata Template

We are utilizing an Internet application, a forms-based Perl script, for describing resources using the Dublin Core standard. It is from Sweden and is available at <http://www.lub.lu.se/cgi-bin/nmdc.pl>. We have plans to build a similar application which would be customized for our information within the next six months.

2.2.4 USGS MetaMaker Tool

We have tested and the software tool recommended by USGS, called MetaMaker (http://www.emtc.nbs.gov/http_data/emtc_spatial/applications/nbiimker.html). MetaMaker is a software tool that is in development. It allows a forms-based data entry to describe data designed to produce a text and HTML version of the data set documentation. Although the EE-IR Center staff has experience using MetaMaker, there are enough problems with exporting data from the software that it is not currently being used. One of the more serious defects that makes the software difficult to use is in its report function, which produces output without all the fields present in the original file. Additionally, the parser that is recommended to be used with this program produces erratic error messages that appear to be false alarms. The serious problems of the MetaMaker is contributing to the slow rate at which the FGDC initiative is growing, for example, in terms of number of metadata records written.

2.2.5 Dublin Core to MARC converter

A crosswalk has already been documented by the Library of Congress for converting Dublin Core to the USMARC standard. We are working with the project software team to build Web based software tool which will convert as a batch multiple Dublin Core records into one USMARC structured file with multiple MARC records, which we will then batch load. We are designing it in such a way that it will also be available for use by others.

2.2.6 Our range of access points

The labor required to write metadata, whether it is Dublin Core or FGDC compliant, is expensive. Software tools for converting those records into other formats, once written, are cheap. The approach we are taking is to convert our metadata into multiple formats, and serve it from multiple access points. Thus, the format of the metadata (Dublin Core, USMARC, or FGDC) determines how it will be indexed, and by extension, what kind of search functionality is possible, and what kind of user will find the metadata.

At the EE-IR Center Web site, the metadata are in Dublin Core records, comprising title list plus search index. Users accessing the EE-IR Center's Web pages can see the Dublin Core records. These are available to be indexed by all the major Internet search engines (AltaVista, Lycos, Excite, etc.), as well as any Internet subject guides.

The EE-IR Center maintains an integrated library automation system called the Cuadra Star system. The Dublin Core Web site and document records will be converted to MARC and placed in this local Cuadra Star system with the existing NWRC library catalog. This system will be searched with a very powerful and refined search interface. It will also be available as a Z39.50 server to other systems.

The USGS will be combining all the catalog records from the regional centers into one catalog, the USGS Union Catalog. Our USMARC records will be sent to this site.

The OCLC InterCAT service is searchable on the Web. After our records are converted to MARC they will be imported into this international database of high quality electronic resources.

The FGDC compliant records we create using the FGDC or NBII standards will be loaded onto a z39.50 server located at the NWRC. This node can then be searched through either the NBII Metadata Server, or the FGDC Clearinghouse.

2.2.7 Future directions

Given its status as the global metadata standard for bibliographic materials, the MARC standard will remain with us for a long time. We consider USMARC our core metadata format.

The future of the FGDC/NBII initiative is less clear, given the problems in the creation of FGDC compliant metadata. There are two main trouble areas. The first is technical. As mentioned above, the bugs in the USGS MetaMaker software tool are so serious as to render the product unfit for use, by our standards. Until this is addressed, the growth rate of metadata production within federal agencies and grant recipients cannot increase. The FGDC group in charge of the initiative appears to be placing their emphasis at this time on making the FGDC standard an international ISO standard. They are also looking for a conversion tool which will

convert the existing FGDC metadata to the emerging ISO standard. Once the ISO standard is in place, they will adopt the ISO standard. The second problem is organizational, and involves the creation of FGDC or NBII compliant metadata. Researchers are reluctant to utilize their time in learning this standard and writing metadata for data sets they collect. Adam Chandler and Dr. Carol Barry at the Louisiana State University School of Library and Information Science are designing a study to better describe and understand this problem.

Within the Dublin Core community, there are efforts taking place which will elevate the DC standard as a viable international general metadata solution. It appears that the Dublin Core standard is heading towards becoming a resource description standard for Extensible Markup Language (XML) documents. That will cause no foreseeable problems for the EE-IR Center, and will in fact add functionality and possibilities.

3 Research and Development of Software Tools

The current and near term goals of the R&D efforts in software tools are: (i) to provide tools and techniques for creating new digital resources and linking them to the Internet; (ii) to acquire/develop tools for defining metadata and building indexes for the resources in the repository; (iii) to acquire/develop advanced tools and techniques for analysis and visualization of data sets pertaining to environmental and energy applications; and (iv) to conduct research on adaptive strategies for multimedia retrieval; knowledge acquisition techniques for multimedia indexing; content-based search for images; methods for mining of data sets via the Web.

Our longer term goals include: (i) to extract and organize expert (e.g., a research librarian) knowledge pertaining to resource identification and search strategies for energy and environmental applications; (ii) to use expert knowledge to build simpler interfaces that enable clients to perform searches on their own; and (iii) to extract and organize expert (e.g., research analysts in the oil and gas industry) knowledge pertaining to the use of tools and techniques for database mining and other related data analysis tasks.

In the following, details of projects that have been implemented and those that are in design phase are described.

3.1 Current or Completed Projects

The following is a list of software projects where we have demonstrable results during the past year.

1. The EGRD (Extraction of Geographically References Data sets) project allows data sets to be extracted from remote URLs based on geographical locations, cleaned, and redisplayed to users in different formats.
2. RUBRIC (Rule Based Retrieval in Computers) allows concept-based retrieval by the construction of a rule tree based on expert knowledge and by providing a software layer on top of some Boolean search engines to convert concept to queries acceptable to the native search engine.
3. The WDB (Web to Database) project enhances an existing general purpose tool for searching a database from the Web.

4. SDMS (Slides Database Management System) is a tool for managing the NWRC Slides Database. Keywords are attached to each slide in the slides database. A user-friendly interface allows users to search and update the slides database.
5. The WBPI (Web to Bayou Periodical Index) project enhances the search capability to an existing Web resource by implementing and evaluating two approaches: structured (relational) database HTML documents enhanced with meta tags.
6. CAST (Catalog Acquisition and Search Tool) is a retrieval engine for searching catalogs on the Internet developed to access multiple catalogs from multiple distributors of various products.
7. The CBIR (Content Based Image Retrieval) system enables images to be retrieved based on their distances to a query image. The system fetches and displays those images which are similar to the input image, based on newly developed techniques for efficiently measuring similarities between images.

3.1.1 Extraction of Geographically Referenced Data sets (EGRD)

The EGRD system is designed to access, filter and display information from different data sets, compiled by third parties and hosted in remote web servers. Extracted information could be saved at our server in a desired format for subsequent downloading and/or displayed in several formats (e.g. tables, graphs, bar charts). According to the user interests he/she can specify a geographical region (e.g. State or Parish) and other criteria (e.g. certain years or certain columns of the data set) in the context of a data set (e.g. Endangered Species) of interest. The software avoids having to store the whole data set (which is typically a very large ASCII file) in the local server. The software provides clickable maps to initially specify geographical regions of interest and uses technologies such as Java applets and Java Remote Method Invocation (RMI). The URL for this project is at <http://mgrss20.cacs.usl.edu/~sxp1258/project/working/index.html>

3.1.2 Rule Based Retrieval in Computers (RUBRIC)

The RUBRIC system is a rule-based information retrieval system designed to provide ranked output by associating Retrieval Status Values(RSVs) to textual documents. A rule-tree is constructed based on an input set of rules. The system allows users to create and/or modify a rule base by providing an interface in which a set of related rules are represented by an AND/OR tree. The edges in the tree can be assigned weights that reflect the degree to which the left hand side of an IF/THEN rule implies the concept in the right hand side. This system can be used both for setting up canned Boolean queries and Concept Based Retrieval. The software is written in a way that it can be added on as a top layer to an existing Boolean Search Engine. Our current efforts are in the direction of demonstrating our ideas in the context of DoE's InfoBridge. The software is implemented using a Java applet. The URL for this project is at <http://www.ucs.usl.edu/~fx17146/rubric/hua.html>

3.1.3 Web to Database System (WDB)

The system is an enhanced version of WDB, a general purpose tool written in Perl for linking tables in a relational database system to the Web and for searching via standard Web

browsers. NWRC has two databases stored in an MiniSQL DB that are provided Web access by means of this software: publications database and water fowl information database. The user-interface has been enhanced to allow more versatile query specifications. Also, some update functions are added. The URL for this project is at http://www.cacs.usl.edu/su98-bin/mke4542/project/nwrc_wdb/bps8527/duck/menu_form

3.1.4 Slides Database Management System (SDMS)

This system manages the slides database of the National Wetland Research Center (NWRC). It is a reengineering of what they used before. Each slide in the database is described by a set of keywords/ formatted attributes. By using these keyword queries, the system allows users to search and update this Database System. Visual Basic language and MS-Access software are used. It has access controls to where scientists can propose changes to descriptions, but only library personnel can make the actual updates.

3.1.5 Web to Bayou Periodical Index (WBPI)

The system enhances the search capabilities of legacy internet resources (e.g. old-style pre-existing resources developed prior to the advent of more modern methods). Two approaches are modeled and implemented. The first approach uses structured databases and the other approach uses HTML documents enhanced with content tags.

Both implementations enable users to specify search terms in particular fields of the resource in order to obtain only that small portion of the resource that match the specified search field values. The URL below is for the approach based on creating a structured DB. We use Pro-C, CGI programs and Oracle DB.

The URL for this project is at <http://www.cacs.usl.edu/cs561-bin/Bayou.cgi>

3.1.6 Catalog Acquisition and Search Tool (CAST)

This is a search engine for electronic product catalogs on the internet. The system is designed and developed to access multiple catalogs from multiple distributors of a given class of products. It is assumed that the critical information that needs to be indexed appears in an HTML document in a tabular form. The needed catalog pages are downloaded to our site and enhanced by the (automatic) insertion of content tags. This downloading is carried out by our own extension of a public domain spider, called the MOM-Spider. Flexible searches of the catalogs are made possible by using a public domain search engine called ISITE, which is capable of exploiting the content tags that have been inserted into the HTML catalogs. We also enhance ISITE with a concept based retrieval capability (using ideas described above in the RUBRIC project). Most of the programming or software modifications we did involves the Perl language. The URL for this project is at <http://www.cacs.usl.edu/~bps8527/CAST/index.html>

3.1.7 Content-based Image Retrieval (CBIR)

An Image Retrieval System, which retrieves a set of images by submitting a query image, is being developed. Depending on the color contents of the query image, the system retrieves those images which are closest to the query image. The goal of this work is to improve the

efficiency of content based image retrieval by using what we call an Estimated Distance, rather than the more expensive Real Distance that users would prefer to adopt. The improvement in efficiency will yield the benefit that the system can deal with an image database consisting of a very large number of images. The current demo shows that, for a chosen small collection of images, the Estimated Distance based on color indeed allows the user to find the images in the "right" order. We use C and CGI programming in this work. The URL for this project is at <http://www.cacs.usl.edu/~xxw4192/CBIR.html>

3.2 Projects in design phase

Several of our projects are in design phase. Their designs are in conjunction with the library group. Two such projects are the design of tools for creating US MARC metadata out of Dublin Core metadata, and for collecting meta data (EGRD). Other projects are as follows.

3.2.1 Data Warehousing tools for Geographically Referenced Data Sets

Data warehouse is a database designed to support decision making in organizations. As scientific databases grow at unprecedented rates, new approaches are necessary to enable scientists to locate and integrate data sets pertinent to their needs. One method that shows promise is the implementation of data marts that integrate appropriate data sets needed for certain subjects (or issues). The goal of this project is to provide a highly integrated tool for decision support over the web for geographically referenced data.

3.2.2 Mining Association Rules Between Sets of Scientific Citations

The goal of this project is to direct users to citations that are highly relevant to the intended request. The project stages are: (i) extract data from a citation database, such as the Web of Science; (ii) load extracted data to a database after transforming them into a different structure; (iii) find all the associations; (iv) present them to the user using visualization tools

3.2.3 Visualization of Multidimensional Relevance Maps

A retrieved object has different levels of relevance to the specified list of features. These relevance levels must be presented to a user in an informative and intuitive manner. Each feature defines a coordinate of the multidimensional information space. However, the features have to be placed judiciously in a 2- or at most 3-dimensional space for a user to be able to understand the relationship among the retrieved objects and features. In our approach, features form gravitational points on a 2- or 3-dimensional space; an object centers around the features that it is most relevant to. We plan to adapt placement algorithms from computer hardware layout design for our purpose.

4 Internet Computing Research

Research in Internet computing is conducted by faculty, staff, and students at CACS. Our research work is organized under the following four headings: data mining, parallel and distributed computing, media technology, and data visualization.

In multimedia retrieval, we have completed a project in description and evaluation of an adaptive image retrieval system called Web-IDBS. We have made a qualitative comparison of different client-server implementations that used the Java programming language for an adaptive image retrieval application. We have developed strategies for improving the efficiency of feature-based (color, texture or shape) image retrieval.

In data mining, we are investigating the impact of data mining on database security. Specifically, we are focusing on evaluating a protected data element's risk of disclosure in the context of classification learning. We partitioned classification methods into two categories, and developed evaluation algorithms for the two categories. We are currently working on experimental investigations of these algorithms. Vijay Raghavan guest edited a special issue of the Journal of American Society for Information Science (April 1998, Wiley) on "Knowledge Discovery and Database Mining". One of the papers in that special issue reports our research results on efficient algorithms for feature selection, in the context of designing classifiers based on the theory of rough sets.

In parallel and distributed processing, we investigated the design of a high-performance server cluster for web applications. A cost effective approach to setting up a web server is to incrementally scale up the hardware as web access traffic increases. Workstation clusters are an attractive solution for such applications. The performance metric in evaluating such solutions is the average waiting time of a task to be completed. As the workload of one workstation reaches a threshold, it can send some of its tasks to its proxies, who are considered to be logical neighbors. We developed such load balancing algorithms for two different network configurations, and evaluated the performances empirically on actual networks in our department. The preliminary results are that the load balancing algorithms provide substantial improvements in average waiting times, especially at high traffic densities.

In media technology, we investigate issues related to transmitting imaging and video data over the Internet. Specifically, we developed a method for progressive transmission of image data so that the rendered image is perceptually lossless. In view of the large number of images that are printed after being downloaded, we incorporated halftoning into the compression method. Video data convey information in the most vivid form, albeit at a cost of substantial bandwidth requirement. For the foreseeable future, the Internet remains a low bit rate channel for video transmission, especially when compared to other channels. We investigated the use of wavelet transform, a method that has proven to be effective for low bit rate image transmission, for low bit rate video compression. Our method used a pair of wavelet-based processing modules as wrappers around an international standard codec (the ITU-T H.263). We showed that we can have better video quality (in terms of still frame and motion) at fixed bit rates compared to the standard codec.

Authentication and intellectual property rights is an important aspect that affects the flow and exchange of information through the Internet. Digital watermarking is the area that studies methods for embedding of authentication marks in data sets. Specifically, our interest is in watermarking image and video data by embedding marks that are not perceptible to a user. Our main concern is with the survival of a watermark through low bit rate compression. Our approach is to modulate a watermark using the luminance channel. We empirically assess the

bit error rate due to compression, and will incorporate error correcting codes that are appropriate for the observed bit error rate into the watermarking process.

In data visualization, we developed a new color halftoning algorithm for indexed color display. Compared to the standard error diffusion algorithm, our new method has fewer color spikes and no color shifts. We also developed a method for visualizing scalar fields on three-dimensional surfaces. The application of our method is in visualizing EEG-derived cortical potential fields on MRI-derived cortical surfaces—a method that fuses two brain imaging modalities.

5 NASA/USL Regional Application Center Enhancement

The mission of RAC is to provide NASA based technologies to the general public; to provide capability to directly and indirectly receive, manipulate and disseminate satellite and other remotely sensed data in real time for application development; and to foster the self-supporting application of environmental and Earth resource data by regional institutions including: federal, state, local government, universities and commercial companies. The objectives of our current activities include promoting public access to NASA data as well as collaboration to refine technology; validating the Mission to Planet Earth (MTPE) data; enabling data fusion; and stimulating related commercial activity. An immediate goal is to integrate current and explore future technologies in ingesting, storing and disseminating data. Additional storage cells are being acquired for the manipulations of raw image ingest data. NT Workstation equipment is being installed for integration into the RAC configuration and development of end user products and applications. We continue to develop mass storage and archival system (over 1 terabyte) for serving imagery, text, and geospatial data.

In the past year, we acquired a GOES geostationary satellite receiver to augment our data collection by the AVHRR receiver. We improved the computing environment by acquiring color workstations and several servers. The NASA/USL RAC is one source of information for our EE-IR Center. One of our ongoing projects involves the development of tools for the integrated use of geographically referenced data sets and for the mining of these data sets to aid in decision making relative to certain biological or environmental issues. The approach involves the re-engineering of applications currently implemented using Geographic Information Systems so that they use state-of-the-art technologies (Java, CORBA, CGI scripts, etc.).

6 Research Publications and Human Resources Development

In the following, we list our publications and presentations that acknowledged this grant. Electronic versions of our publications are available from the URL: <http://eeirc.nwrc.gov/pubs.htm>. The titles of dissertations, theses, and projects by students who received support from this grant are also listed.

6.1 Publications

6.1.1 Journal Articles

1. S. K. Choubey and V. V. Raghavan, "Generic and Fully Automatic Content-based Image Retrieval Using Color," *Pattern Recognition Letters*, vol. 18, nos. 11-13, Nov. 1997, pp. 1233-1240.
2. S. K. Choubey and J. S. Deogun and V. V. Raghavan and H. Sever, "On Feature Selection and Effective Classifiers," *J. Amer. Soc. for Information Sci.*, vol. 49, no. 4, April 1998, pp. 423-434.
3. V. V. Raghavan and J. S. Deogun and H. Sever, "Data Mining: Trends and Issues—Guest Editors' Introduction," *J. Amer. Soc. for Information Sci.*, vol. 49, no. 4, April 1998, pp. 397-402.
4. P. Bollmann-Sdorra and V. V. Raghavan, "On the Necessity of Term Dependence in a Query Space for Weighted Retrieval," *J. Amer. Soc. for Information Sci.*, vol. 49, no. 8, August 1998.

6.1.2 Conference Articles

1. H. Sever and V. V. Raghavan and T. D. Johnsten, "The Status of Research on Rough Sets for Knowledge Discovery in Databases," *ICNPAA-98: Second Int'l Conf. On Nonlinear Problems in Aviation and Aerospace*, 29 April - 1 May 1998, Daytona Beach, FL.
2. M. C. Erie and S. M. LeBlanc and V. V. Raghavan, "Enhancing Search Capabilities of Legacy Internet Resources," *InForum 98- Science at the Desktop: Synergy through Sharing*, 6-7 May 98, Oak Ridge, TN.
3. J. S. Deogun, V. V. Raghavan, and H. Sever, "Association mining and formal concept analysis," *Proc. of RSDMGrC98—Sixth International Workshop on Rough Sets, Data Mining and Granular Computing*, Research Triangle Park, NC, Oct. 1998.
4. J. S. Deogun and H. Sever and V. V. Raghavan, "Structural Abstractions of Hypertext Documents for Web-based Retrieval", *DEXA 98—9th International Conference on Database and Expert Systems Applications*, 24-28 August 1998, Vienna, Austria.
5. M. C. Erie and C. H. Chu and R. D. Sidman, "Visualization of Reconstructed Potential Field on MRI-Derived Scalp and Cortical Surfaces", *American Clinical Neurophysiological Society Annual Meeting*, 3-5 October 1998, New Orleans, LA.

6.2 Graduate Student Dissertations, Projects, and Theses

1. Nancy Breaux, "Data compression and halftone rendering for gray scale and color images," Ph.D. (Computer Engineering) Dissertation, May 1998.
2. Wei-Kian Chen, "Re-engineering of the Belfast newsletter index database system for efficient access via the Internet," M.S. (Computer Science) Thesis, May 1998.
3. Thanee Dechsakulthorn, "Design of a high-performance server cluster for web applications," M.S. (Computer Engineering) Project, May 1998.
4. Hao Duan, "Wavelet transform-based methods for low bit rate video compression," Ph.D. (Computer Engineering) Dissertation, May 1998.

5. Thomas Johnsten, “Impact of data mining on database security,” Ph.D. (Computer Science) Dissertation, August 1998.
6. Shiyong Ning, “Description and evaluation of Web-IDBS—An adaptive image retrieval system,” M.S. (Computer Science) Project, May 1998.
7. Balaji Parthasarathy, “Qualitative comparison of different client-server implementations of an adaptive image retrieval application in Java,” M.S. (Computer Science) Project, Dec. 1997.
8. Bhanu Suravarapu, “An intermediary-based approach for searching product catalogs on the Internet,” M.S. (Computer Science) Thesis, May 1998.