

**UNIVERSITY OF SOUTHWESTERN
LOUISIANA**

**ENERGY AND ENVIRONMENTAL
TECHNOLOGY APPLICATIONS PROGRAM
(EETAP)**

RESEARCH COMPONENT 1

**"Information Systems Technology for Energy and
Environmental Applications"**

Dr. Vijay Raghavan, Principal Investigator

Revised

May 16, 1997

TABLE OF CONTENTS

BUDGET	i
BUDGET EXPLANATION	v
COST SHARING	xi
1. INTRODUCTION	1
1.1 Motivation	1
1.2 Specific Objectives And Significance Of The Expected Results	2
1.3 Organization of Narrative	2
2. INFORMATION CENTER DEVELOPMENT	3
2.1 Relation To Present State Of Knowledge	3
2.1.1 Digital Libraries	3
2.1.2 Database Mining	4
2.2 Development Plans	5
2.2.1 Plans For Services And Resources	5
2.2.2 Potential Clients	6
2.2.3 Types Of Information Materials And Data	6
2.2.4 Information Center Development	7
2.3 Technical Approach	9
2.3.1 Metadata Format	9
2.3.2 Protocol-level Interoperability	10
2.3.3 Data Format	11
2.3.4 Work-centered Information Services	11
2.3.5 Semantic Interoperability	12
3. RESEARCH PROGRAM	12
3.1 Adaptive Multimedia Retrieval	12
3.2 Multimedia Indexing and Personal Construct Theory	15
3.3 Feature-based Image Retrieval	17
3.4 Database Mining	20
3.4.1 Rough Sets	20
3.4.2 WebMine	20
3.5 Application Partitioning for Distributed Computation	21

3.5.1 Data Replication	24
3.5.2 Run-time data communication	24
3.5.3 Optimizing number of processors and the message size	25
4. REGIONAL VALIDATION CENTER ENHANCEMENT	27
4.1 Background	27
4.2 RVC Concept	27
4.3 RVC Initial System Functionality	27
4.3.1 Ingest, catalog, and store satellite data	28
4.3.2 Support registration of algorithms, schemes, and resources	29
4.3.3 Support database queries through a graphical user interface	29
4.3.4 Integrate planning and scheduling capabilities	29
4.4 Rationale for Enhancement	30
5. SCHEDULE OF WORK	32
6. BIBLIOGRAPHY	33
7. BIOGRAPHICAL SKETCHES	36
8. CURRENT AND PENDING SUPPORT	57
9. APPENDIX: LETTERS OF SUPPORT	58

1. INTRODUCTION

Energy demand from oil and gas is expected to remain at high levels well into the future. The Gulf of Mexico region is the source of almost 85% of the nation's oil and natural gas (including imports). It is a region rich in natural resources, but it is also an environmentally fragile area, subject to floods, hurricanes, and potential oil spills. Prudent development of public and private energy and natural resources depends on having information and technologies to understand, quantify, and assess the risks and consequences of oil and gas development and transportation. Besides being the major source of oil and gas, the Gulf of Mexico has a significant link to the nation's economy. Gulf ports handle over half of the nation's import-export tonnage, while five of the top-ten fishing ports are located on the Gulf. Timely and ready access to relevant information, especially by the energy industries, is critical for environmentally sound utilization of this ecological system. The University of Southwestern Louisiana (USL), in collaboration with the National Wetlands Research Center (NWRC), therefore proposes to build on their scientific expertise and facilities to develop state of the art digital information and analysis technologies to support research related to energy and environmental issues.

The project centers around the development of a comprehensive, energy and environmental information center, based on software agents, to create metadata to locate appropriate information sources, and to provide users with transparent searches of the relevant databases. The development of this Information Center will build on an existing regional scientific data center. To support the Information Center, the data center will be enhanced to provide capabilities for data ingest from a wide variety of sources; for calibration, registration, and storage; for real-time fusion of data sets from multiple sources; and for efficient display and effective user interface.

The Principal Investigators at USL's Center for Advanced Computer Studies (CACS) will work with information professionals of the National Wetlands Research Center—a part of the U.S. Geological Survey's Biological Resources Division—to create the Information Center. This center would create or procure enabling technologies and develop human resources that allow ease of finding and interpreting information needed by energy and environmental applications. This Information Center will benefit users by capturing and organizing the knowledge and experience of environmental research librarians, by creating metadata to locate appropriate information sources, by providing software tools and infrastructure to transparently (without having to learn a variety of search engines and user interfaces) search the databases, and by creating a local repository of data.

1.1 Motivation

We are indeed in the midst of the information age as evidenced by the phenomenal growth in the acceptability by the general public of the Internet (the Net)—in particular, of the World-Wide Web (the Web)—as the source individuals turn to for information in many areas. This distributed collection of information resources and services has the *potential* to be enormously helpful in performing information-intensive tasks, such as those found in

environmental and energy applications. But these information sources are extremely varied in content and format (e.g., data sets of biological information; enormous amounts of weather and climate data; continuous data logger information; research data; species longevity data; sociological data; bibliographic data; spatial data; metadata about biological and spatial data). Accessing available resources is further compounded by the problem of variability in client/server protocols. The Dialog information services, for example, is a Telnet-based information provider, while the WebCrawler, which indexes documents on the Web and returns their URLs in response to queries, is an HTTP-based service. Furthermore, there is practically no evaluation of the usefulness or reliability of the data available on the Net. Because of such reasons, information intensive-tasks are turning, more often than not, into confusing, frustrating annoyances by forcing programmers, librarians and users to learn many interfaces and by confronting end-users with bewildering details of fee-based services.

Another challenge in harnessing available information resources and library services is the understanding and addressing of one's client base. While several research groups are looking into needs of education- or entertainment-related applications (e.g. digital library services for high school students), those solutions do not adequately meet the specialized needs of other user groups. In response, a novel notion called work-centered digital information systems has come into vogue. The potential clients of the Information Center will include analysts and researchers concerned with energy and environmental issues, whose needs are at present largely unmet.

1.2 Specific Objectives And Significance Of The Expected Results

The objectives of this work is to develop, in a three-year period, an Information Center that focuses on energy and environmental resources. The Information Center will develop a repository of information, especially those pertaining to the Gulf Coast region, that are important for energy and environmental operations. The Information Center will also develop software tools based on the enabling technologies of digital library and data mining. These software tools can be considered as agents in the sense that an agent meets the requests of a user by providing the relevant information and, depending on the nature of the requests and search results, suggests analysis tools.

The results will be significant in two main aspects. First, the Information Center will be a model of applications of digital library and data mining technologies, especially in the areas of geographically referenced data. Second, the repository materials will be organized and be available for decision makers, research scientists, and others who are interested in information especially around the Gulf coast.

1.3 Organization of Narrative

There are three components of work for the creation of the Information Center: (a) the basic structure of the center; (b) research projects that support the development of advanced tools used in the center; and (c) enhancement of the regional scientific data center (the USL RVC) for stocking the center with a local repository of data. Details of work involved in these

three components are respectively described in Sections 2, 3, and 4. The work schedule is described in Section 5.

2. Information Center Development

2.1 RELATION TO PRESENT STATE OF KNOWLEDGE

In addition to having a research librarian who is familiar with environmental and energy databases and can evaluate information and provide mediated searches, two information technology solutions can be useful for the clients to retrieve disparate data. First, software tools—such as ones based on information retrieval, expert systems, or pattern recognition—would help users without professional information science backgrounds to perform sophisticated searches easily and gain access to databases in varying formats and needing various search strategies. Secondly, advanced analysis software would combine databases and detect patterns and trends in the data. Therefore, trends in these two emerging areas of information technology, namely, digital libraries and database mining, are of paramount importance to the development of the proposed Information Center. In the following, we summarize the recent results in digital libraries and database mining that can form the building blocks of the Information Center.

2.1.1 Digital Libraries

Digital library research is primarily concerned with providing user-friendly, efficient search capability for users to gain access to relevant digital information that can come from disparate sources. In this era of the Internet and the Web, the long time topic of digital libraries has again become white hot. As more and more people get connected to the Net and the distributed repositories grow in size, the need for effective and efficient tools has become critical. As the *Science* news articles on the U.S. Digital Library Initiative (DLI) have put it, the Internet is like a gigantic library without a card catalog [Pool94, Chan95]. At present even the most sophisticated search engines, such as Altavista, Yahoo, and Lycos, are not wholly adequate to provide real ease of access to the mass of information available.

This is why the U.S. government made digital libraries the flagship research effort for the National Information Infrastructure (NII), which seeks to bring the highways of knowledge to every American. Six research groups are involved in the DLI, which is sponsored by the National Science Foundation (NSF), the National Aeronautics and Space Administration (NASA) and the Advanced Research Projects Agency. The six lead universities are: the University of Illinois at Urbana, the University of California at Berkeley, the University of California at Santa Barbara, Carnegie Mellon University, Stanford University and the University of Michigan. The four-year, multi-agency DLI was funded with roughly \$1 million per year for each project. These Digital Library research projects have a common theme of bringing user-friendly and meaningful search capability to the ever-expanding Net.

While all DLI projects have some connection to this proposal, the goals and plans of some of them are particularly relevant. First, the emphasis of the University of California at Santa Barbara's project, called *Alexandria Digital Library (ADL)*, is on geographically referenced

materials [Smit96]. We will therefore be able to build on the research results developed by ADL, especially with respect to the types of repository and metadata formats used in the software tools. Secondly, issues of interoperability among disparate, existing services and resources are studied by Stanford University, the University of Illinois at Urbana and the University of Michigan. The Stanford approach emphasizes interoperability between protocols by means of translators [Paep96], while the University of Illinois takes the federated approach, based on SGML tags, to minimize variability in formats [Scha96]. The University of Michigan group is exploring an approach based on interacting software agents that cooperate (and compete), in order to meet the information needs of their clients [Atki96]. In particular, there are agent types such as user interface agents, mediator agents and collection interface agents. Thirdly, the University of California at Berkeley project is creating a prototype set of information services called the California Environmental Digital Information System, which includes many different kinds of environmental data [Wile96]. Their results on work-centered information services are of particular interest to our project.

2.1.2 Database Mining

Advanced software tools can often assist a user in analyzing search results by, for instance, detecting patterns and trends in the large amount of data. In the last decade, we have seen an explosive growth in our capabilities to both collect, retrieve, and generate data. Examples of this growth can be found in all arenas such as scientific, business and governmental data. Our inability to interpret and digest these data, as readily as they are accumulated, has created a need for a new generation of tools and techniques for automated and intelligent database analysis. The study of such tools and techniques is the subject of the rapidly emerging field of Knowledge Discovery in Databases (KDD) [Fayy96]. KDD and other phrases, such as database mining, information harvesting or data mining, have been used to refer to the notion of finding useful patterns (or nuggets of knowledge) in the raw data.

KDD systems typically draw upon methods from diverse fields such as machine learning, pattern recognition, database management, statistics, knowledge acquisition for expert systems, and data visualization. Research results from the areas of pattern recognition and machine learning are relevant in the sense that they provide the theories and algorithms for systems that extract patterns and models from data. However, KDD research enables the application of these theories and models to large data sets. KDD also has much in common with statistics and exploratory data analysis, particularly in terms of statistical procedures for modeling data and handling noise. Machine discovery, which targets automated discovery of empirical laws from observation and experimentation, is a closely related area. In business environments, the notion of data warehousing, which refers to the recently popular trend of collecting and cleaning transactional data and making them available for on-line retrieval, is becoming popular. KDD and On-line analytical processing (OLAP), intended for analyzing data warehouses, are related in that they both share the goal of providing a new generation of intelligent information extraction and management tools. Machine learning and knowledge acquisition in expert systems are clearly related, except

that knowledge is extracted in expert systems through the interactions of a knowledge engineer with an expert in the application domain.

The emphasis of our project is on work-centered digital library services, with plans to provide our clients with tools and techniques for further analysis of data sets found through searching digital library resources. This choice emphasizes the inclusion of database mining tools within the scope of services and resources offered by the Information Center.

2.2 DEVELOPMENT PLANS

2.2.1 Plans For Services And Resources

In general, the plans call for establishing a repository of resources, such as those in the Gulf Coast region, that are currently not available in digitized form, building an infrastructure that makes digital resources that are relevant to the needs of the clients more easily accessible, and providing services that enable analysis and interpretation of search results. More specifically, the planned activities include:

- Prioritize all resources, whether geographically local or remote, that are of interest to the client-base.
- Create a digital library testbed consisting of documents and data sets that are either not in digital form or are not yet organized for ready access.
- Acquire/develop tools for defining metadata and building indexes for the resources in the repository.
- Provide tools and techniques for linking new digital resources to the Net.
- Conduct searches based on request for references or data sets.
- Extract and organize expert (e.g., a research librarian) knowledge pertaining to resource identification and search strategies for energy and environmental applications.
- Use expert knowledge to build simpler interfaces that enable clients to perform searches on their own.
- Hold workshops to provide training and assistance to clients interested in establishing testbeds, creating metadata, and linking them to the Net, and to refine the research agenda.
- Acquire/develop advanced tools and techniques for analysis and interpretation of data sets pertaining to environmental and energy applications.
- Perform state-of-the-art analysis and interpretation of data sets as requested by clients.
- 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.
- Hold workshops to provide training and assistance to clients interested in analysis, interpretation and visualization of data sets, and to refine the research agenda.
- Provide support for data mining and other interpretation tasks by using parallel computing technology.

- Conduct research on adaptive strategies for multi-media retrieval; knowledge acquisition techniques for multimedia indexing; content-based search by color, texture and shape for images; and, methods for mining of data sets via the Web.

2.2.2 Potential Clients

The primary clients of the Information Center would be those with interest in the Gulf of Mexico, including oil companies, private consulting companies, non-governmental organizations, government agencies at various levels, attorneys, construction companies, shipbuilders, coastal developers, faculty and students, and the general public. Generally, the clients would be those who lack their own information professionals, such as research librarians, computer systems professionals and data analysts, but have access to a computer so that they could use digital information sources and services. Primary clients ideally would be able to pay for the service at a reasonable cost-recovery rate.

We will establish an expert Advisory Board, consisting of representatives from the Department of Energy, scientists from the NWRC and other agencies, and corporations that are representative of users who are interested in energy and environmental information. The Advisory Board will meet quarterly, and will provide feedback and guide our efforts from a user's perspective.

2.2.3 Types Of Information Materials And Data

Information that will be handled by the Information Center would run the gamut from paper documents in "grey" government literature to spatial and image data. Based on currently available library resources at the NWRC and the Net resources typically accessed by the research librarian at NWRC, we identified the following resources that are of potential interest, in terms of the nature and types, to our clients. Though the list is not complete, it is adequate for the purpose of developing our approach. We expect to further refine this information after consultation with our Advisory Board.

- Data sets in known and unknown digital formats and in different media (paper, CD-ROM, tape, diskette) that would need to be converted to appropriate format for processing by database management software, spreadsheet software, or GIS (geographic information software) for spatial data visualization. Examples are data sets of plant and animal names (taxonomic) by geographic regions, water levels at specific stations, geographic digital line graphs, digitized coastlines, digital topographic relief images, digital bathymetry at various scales, ocean-drilling data, seismic profiles, geochemical database on offshore mineral resources, hydrographic survey data, gravity data for the United States, earthquake data, Permanent Service for Mean Sea Level (PSML) sea level data, ocean temperature and salinity, wind speeds, rainfall total and ranges, temperature extremes, solar radiation, and historical hurricane data.
- Data directories of biological data such as the U.S. Geological Survey's National Biological Information Infrastructure (NBII). The NBII contains metadata and direct access to some data sets in the directory held primarily by centers in the USGS

Biological Resources Division (BRD). The Information Center will both use the directory resources and contribute metadata to populate the directory. Biological data in the NBII metadata will include data from any agency as the directory grows.

- National Spatial Data Infrastructure (NSDI), a network of spatial metadata and data required by federal mandate for all federal data. The NWRC is one of the prime leaders in the NSDI through training in writing metadata standards and producing prototype systems for serving the data. The BRD has designed systems to automate metadata. The Information Center will add metadata to the USGS node of the NSDI.
- Digital images stored in formats such as the GIF or TIFF. Examples include maps; slides of flora or fauna, habitat, coastal land use; videos from the local collection; image data retrieved from the Internet; and databases of slides with textual information. For example, the NWRC is developing a large image database of coastal subjects, exotic plant species, and forest species.
- Sources, pathways, and results of database searches in local or online databases to retrieve, for example, coordinates for a geographic location, addresses and phone numbers of experts, taxonomy of flora or fauna, publication metadata, definitions, foreign words and phrases, weather stations, hydrological station. Many different databases may need to be searched to get one piece of information.
- Raw data from continuous monitoring of the environment such as NASA weather data, global position systems, or data loggers in the field.
- Text documents in various digitized formats such as Portable Document Format, WordPerfect, Word or plain ASCII text. Examples of documents include journal articles, journal table of contents, entire publications, federal regulations, text of bills pending in U.S. Congress or in state legislatures (if available), U.S. Code sections, bibliographic data, metadata for publications, names of libraries holding a specific publication or journal, publication lists, and Supreme Court dockets and decisions.
- Paper copies of hard-to-find technical reports, environmental impact statements, file reports, unpublished reports written by government agencies or by consultants and non-governmental agencies.

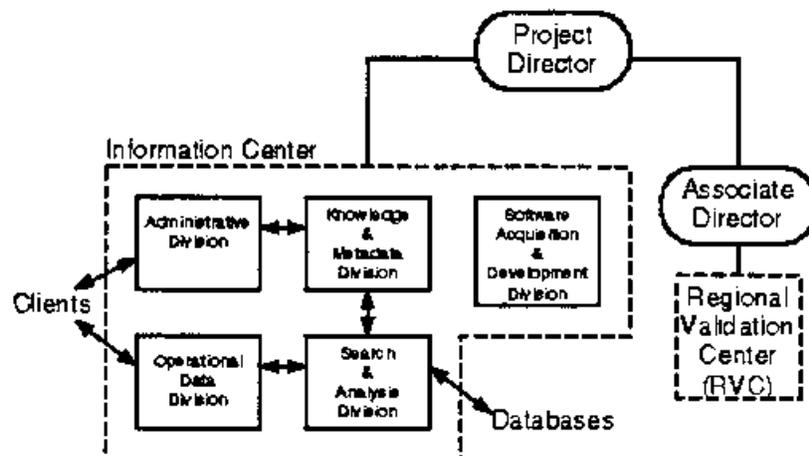
2.2.4 Information Center Development

The organizational structure of the Information Center is as follows:

◊ *Knowledge & Metadata Division*

Creates local metadata databases, generates indexes/catalogs for repository data, manages metadata, builds predefined queries and enables query reuse, performs knowledge engineering tasks for work-centered information services.

- ◇ *Search & Analysis Division*
Chooses appropriate sources; selects search plans; performs conversion, retrieval and updating; Applies analysis, data mining, and visualization tools.
- ◇ *Operational Data Division*
Makes quality assessment of data sources, such as NASA, NOAA, USGS; creates local database through special-purpose ingest software, authoring tools, scanners, etc. Provides services such as creating CD-ROMs for distribution of results/data products.
- ◇ *Software Acquisition & Development Division*
End-user interfaces, report writers, database mining tools, adaptive retrieval software. Commercial tools will be acquired as necessary; other tools will be developed by the investigators.
- ◇ *Administrative Division*
Interfaces with clients; handles security issues, chargeback, client request tracking, resource allocation, copyright permissions, etc.



There will be a considerable amount of repository materials and information that will be in the public domain. Other interactions might involve user-fees, depending on the sources of the information; such an interaction between different divisions can be illustrated by the following example:

When a client makes a request to the Information Center, the Administrative Division checks the nature of the request, determines processing schedule, allocates resources and processes a chargeback, if applicable. Suppose the request is for a complex search involving several data sources; the Knowledge & Metadata Division devises a search plan, e.g., that a predefined query be used to locate tide data at NOAA, and that regionally collected data be considered for comparison. The Search & Analysis Division conducts the search, and prepares a list of references to the data set(s). The Operational Data Division would note that the tide data are in comma delimited format and, since it already has metadata, that no conversion is needed and the data can be incorporated into a spreadsheet for the client.

The Information Center will be developed based on the following plan.

- Working with potential public and private clients, a list of the most needed kinds of information will be developed. For example, a list of the web sites accessed most frequently in the past six months by the NSW research librarian includes:
 - ⇒ Tide predictor- <http://tbone.biol.sc.edu/tide/sitesel.html>
 - ⇒ Historical and current environmental data--temperature, precipitation, cloud cover, water temperature, salinity, water level at the NOAA environmental data directory- <http://www.esdim.noaa.gov/NOAA-Catalog/full-text.html>
 - ⇒ Directory of Biological Data for the United States, primarily data held by the Biological Resources Division of the U.S. Geological Survey- <http://www.nbs.gov/nbii/directory.html>
 - ⇒ USGS Node of the National Geospatial Data Clearinghouse. This site is a searchable data base for georeferenced data- <http://nsdi.usgs.gov/nsdi/>
- The complete list is available at the URL <http://extreme.cacs.usl.edu/~cice/sites.html>.
- Working from this list, a survey will be made of the kinds of data and their value to answering resource/ analysis requests.
- For data in this list, metadata will be written for those data sets currently lacking metadata using the NBII standard for biological data (and related statistical techniques) and the NSDI for spatial data.
- Graphical user interfaces, through which complex queries for data covering a variety of formats and sources listed above can be entered, will be developed. Special software tools to support interoperability among various resources will be provided. The system will allow display of metadata and full text of indexed documents; access to table of contents of indexed documents; and display images. It will utilize commercial application software such as spreadsheets, geographic information systems, statistical analysis software, and graphics packages for data manipulation and visualization; will store information seeking paths for potential repeated queries; and will organize information for future queries.
- Sophisticated information retrieval and analysis tools, such as those for adaptive retrieval and data mining, will be developed to support the databases and clients' needs.

2.3 TECHNICAL APPROACH

In this subsection we address a number of technical issues that will drive our approach to the development of the Information Center. The following is a list of issues that is critical to the success of this effort: metadata format, protocol-level interoperability, (text) data format, work-centered information services and semantic interoperability.

2.3.1 Metadata Format

In regards to the structure of metadata about documents and repositories, our context is closest to that of the University of California at Santa Barbara's ADL project. ADL's experience shows that a cataloging model suitable for traditional library materials (based on author, title and subject) is inadequate for geographically referenced holdings, such as maps and images (of our natural environment). Their goals have been to extend catalog and metadata models and support interoperability between competing models for representing catalog information.

To meet these goals, the ADL project combines the U.S. Machine Readable Cataloging (USMARC) standard and the Federal Geographic Data Committee (FGDC) metadata standard [FGDC1994]. The USMARC standard can be used for storing library holdings' metadata, while the FGDC standard is useful for cataloging digital geospatial metadata, especially since it allows the specification of the relation of different fields within a hierarchical structure. With the combined representation, ADL has been able to catalog all forms of spatial data, including remote-sensed imagery, digitized maps, digitized raster and vector data sets, text, videos and remote Web servers. Although ADL's external interface is via the Web protocol (HTTP), the ADL catalog supports a Z39.5 interface. This is significant since Z39.5 is the NSDI's current search protocol standard. Given the nature of information resources for energy and environmental applications, we anticipate adopting ADL's combined catalog model.

2.3.2 Protocol-level Interoperability

Standardization of catalog and metadata models is but one aspect of interoperability. The other critical issue has to do with protocols for the effective use of client-server models. For the situation where information sources are adopting different protocols, Stanford project has developed a protocol in cooperation with researchers at the University of Illinois at Urbana, the University of Michigan and the University of California at Santa Barbara. They have used CORBA, the distributed-object standard developed by the Object Management Group, to implement information access and payment protocols. Stanford group's implementation is based on that of Xerox PARC's ILU (InterLanguage Unification) facility, a public domain implementation of CORBA. Their approach involves the development of a, so called, multisearch program which accepts a query and multiplexes it to several information sources. The programmer of multisearch program has a common, abstract view of the protocol. The mapping from abstract protocol to those of remote sites is taken care of by *translator* modules. The Stanford protocol is more flexible than existing ones, including Z39.5.

The University of Michigan group chooses a different approach to the implementation of interoperability among protocols. They take the view that none of the current protocols may eventually become *the* standard and opt for the use of *distributed-agent* architecture. Agents are specialized, fine-grained software modules that promote modularity, flexibility and incrementality. In contrast to the use of translators, agent technology requires *mediator* modules between each pair of information source and user interface. In fact, three agent types populate the University of Michigan Digital Library (UMDL): User interface agents (UIAs) manage the interface between human users and UMDL resources; Mediator agents provide intermediate information services and may deal, exclusively, other agents (rather than end users or collections); Collection interface agents (CIAs) act as interface to collections, which are defined bodies of library content. Typically, agent teams are needed to perform complex tasks.

There are other protocol options, such as ZQL, which is proposed by the University of California at Berkeley and whose name derives from its relationship to the Structured Query Language (SQL) and the Z39.5 protocols. Before a particular choice is made, for the Information Center, with respect to protocol-level interoperability, available options such as those mentioned above need to be investigated further. In particular, the protocols used by information resources frequently used by NWRC researchers should be taken into account.

2.3.3 Data Format

For text collections, which may include text, figures, tables, images and mathematical equations, a robust text standard is a must. The University of Illinois at Urbana has decided to choose the Standard Generalized Markup Language (SGML), foregoing, for example, the Portable Document Format (PDF), the HyperText Markup Language (HTML) and ASCII. This has the advantage that many book and journal publishers have already opted for SGML. SGML's strength, in terms of retrieval, is that it reveals deep document structure. It is also possible to readily acquire software such as Open Text's Open Text Index search engine, for indexing and accessing SGML document collections, and SoftQuad's Panorama SGML viewer. For new text collections that will be created at the Information Center, SGML format should be seriously considered.

2.3.4 Work-centered Information Services

Work-centered information services are intended to meet special needs of work-groups. First, work groups are characterized by their desire to retrieve information, rather than documents *per se*. Answering their query may require complex and powerful analysis of selected items. Second, they may want to continually create its own collections of varying types, besides using publicly available resources. Work-groups may differ in their work practices, and may require specialized interfaces as well as system interoperability between analysis and search tools.

Our efforts towards work-centered services, will consider the needs of our clients from the point of view of mining the data sets selected through the searching of various data repositories. In particular, we want to achieve interoperability between special analysis tools (e.g. various computer models, such as those used for predicting effects on vegetation from various hydrologic alterations; decision-support systems, for evaluating coastal restoration needs, or for effects of hurricane on vegetation) and search tools available for searching digital information sources. The research activities described in the *Research Program* section, under the headings of feature-based image retrieval and database mining will provide the basic results needed for the analysis tools to be developed at the Information Center.

We also plan to provide several different user interface alternatives. For example, interfaces for research librarians may be quite different from those directly used by the Information Center clients. The information science literature shows that providing different search interfaces tuned to each search need (type) helps users to become adept in finding information. We plan to conduct such experiments with our user population.

2.3.5 Semantic Interoperability

The kinds of issues considered in the earlier paragraphs mainly emphasize providing access to raw data by achieving syntactic and protocol interoperability. But the goal of accessing high quality, useful information is still elusive. To reach this goal one has to address problems of semantic interoperability. The recent report, on the research agenda for digital libraries, from a workshop sponsored by the Information Infrastructure Technology and Applications (IITA) committee (the primary technical committee for setting National Information Infrastructure (NII) directions for federal government R&D investment) states that “deep semantic interoperability is the *Grand Challenge* for digital library research” [IITA95].

There are many approaches being pursued by the research community to address this challenge. One of the goals is to deal with semantic interoperability issues by having mechanisms in the system to adapt to users based on their feedback. Activities in this direction are provided in the next section, under the title of adaptive multimedia retrieval. We have developed a computer-aided approach to elicit “semantic attributes” from subject experts. This approach is outlined in the next section, under the title of multimedia indexing and personal construct theory.

Semantic attributes allow documents to be indexed by *concepts*. Once the concepts applicable to a collection of documents are known, they can be related to syntactical features of documents by means of rules. This could be accomplished through a knowledge engineering process or, automatically, using machine learning and/or pattern recognition techniques. Our experience with automatic concept learning algorithms is outlined in the next section, under the headings of adaptive multimedia retrieval and database mining. Furthermore, the collection of rules can be organized into a *concept space* that reflects the inter-relationships among concepts. It is known that concept spaces lead to an approach for semantic federation across digital repositories. Recently, Dr. Raghavan, jointly with researchers at University of Nebraska- Lincoln, has recently submitted a proposal to the NSF entitled, “Formal Concept Analysis and Its Applications to Information Retrieval.” Results from this project will also contribute to the techniques and software tools to be developed at the Information Center.

3. RESEARCH PROGRAM

3.1 Adaptive Multimedia Retrieval

In an information-based work environment, where information is acquired in heterogeneous formats, tools for ranking-based, efficient, and customized access to multimedia information are essential to improve the work productivity. The main focus of our effort will be on innovative ranking strategies that are intended to be used in applications that

- i. involve heterogeneous objects such as video, images, sound, text and formatted data,
- ii. need customization of ranking based upon user preferences,
- iii. involve variability in user preference structures, and
- iv. require the distribution of ranking software between client and server processes.

In essence, the kinds of ranking strategies sought will advance the technology of information access much beyond what is possible with traditional querying techniques, such as SQL. These strategies must be capable of discovering patterns in the data in order that retrieval can be performed either by deriving a rule involving a combination of properties that distinguishes objects in a class of interest from objects in other classes (classification), or by finding properties common to objects of the class of interest (characterization). In order to meet such requirements, the proposed ranking strategies will have innovation in the following respects:

- i. The properties describing objects can be keywords, attribute-value pairs, picture subsets, frequency histograms, etc. This will result in the development of generic software components for the retrieval of objects from video, sound, text, and image databases.
- ii. The rank ordering will depend, not only on the distributional characteristics of object properties in the whole database, but also on the preference (or, relevance) information supplied by users, during a search session or with respect to previous searches.
- iii. In order to customize ranking, in addition to exploring preference information supplied by user, preference information may be provided using a two-level (relevant or non-relevant) or a multi-level preference structure.
- iv. The responsibility for ranking will be distributed between client and server processes in an optimal fashion.

An initial study of up-to-date literature revealed that much of the prior effort has emphasized retrieval of information from text databases. In particular, approaches known for ranking of retrieval system output have been developed, for the most part, in the context of textual objects. However, an important requirement in many search and retrieval applications is the capability to accommodate object heterogeneity. One characteristic of text is that they have a natural representation as a set of keywords or a vector of weights associated with the keywords. Such a representation is assumed by most of the ranking mechanisms. In other words, ranking algorithms require that each object be represented as feature vector. In addition to this, ranking algorithms also expect some uniformity as far as the types of attributes are concerned. For example, it may be assumed that the values of a feature are binary (0 or 1) or that they are from the set of natural numbers ($\{0, 1, 2, \dots\}$). It is further assumed that all features have the same domain.

When attempting to devise ranking algorithms for non-textual objects, the assumptions stated above represent a barrier. In traditional formatted databases, for example, a particular object is represented by several types of attributes. Some attributes may be binary valued while others are nominal (e.g., Eye Color), ordinal (e.g., Height \in {tall, medium, short}), from the domain of real numbers, etc. In the case of photographs, videos, etc., the situation can be even more complex. For example, the attribute values may be a complex type such as 2-D string[Chan89] or a frequency histogram. We call such attributes as non-alphanumeric. Thus, two important issues here are:

- i. do we devise entirely new ranking algorithms for non-textual objects, or do we somehow transform non-alphanumeric properties into alphanumeric features?
- ii. even if all features have alphanumeric values, how do we cope with the assumption, of ranking algorithms, that the domains of all features are identical?

Another important research direction we consider is related to the restriction that most query reformulation algorithms impose on the nature of user feedback. In text retrieval, typically the user feedback indicates whether a retrieved document is relevant or not. This corresponds to the assumption that the user has a two-level preference structure. In this context, the goal of ranking algorithms is to ensure that relevant documents are placed higher in the ranked order than the non-relevant ones. In other words, the relative ordering of relevant (or non-relevant) documents is not considered to be of any importance. More generally, however, a user may want to use different degrees of relevance. This situation corresponds to multi-level preference structure. In this case, the ranking algorithm is expected to obtain a ranking such that a preferred object will be assigned a higher position in the rank order than a less preferred one. Since objects similar to preferred ones will also get higher ranks, this kind of search enables a user to guide the search in the direction of his/her interests. The problem of finding an optimal ranking is more general than the 2-class classification problem, but is less general than the multi-class classification problem traditionally studied in pattern recognition.

A third important research direction we plan to investigate concerns the kind of data transfer and communication protocol that is needed between client and server processes in order to carry out the ranking task. In the ranking algorithms currently employed in systems based on client/server architecture such as WAIS [Obra93], the ranking of documents with respect to an initial query is done at the server process since all the distributional information about keywords is available there. The re-formulation of queries based on relevance feedback is either not implemented or is overly simplified. But the kinds of ranking algorithms we envisage will involve more exchange of data and control information between server and client processes. In other words, the distribution of responsibility for generating rank order should be decided upon based on the amount of data transfer that is needed and the number of clients expected to access a server simultaneously. Our investigation of this issue will lead to optimal ways of distributing query reformulation algorithms between browser (client) and server nodes of the World Wide Web [Obra93].

We propose to provide innovative results in all the research directions identified above. In what follows, we outline the approach we will follow in each of these directions and provide a list of associated tasks.

For the problem of dealing with the heterogeneous objects, our approach will be to first demonstrate that certain algorithms currently available to obtain an optimal ranking, for data in which all attributes are derived from the same domain, can be cleverly modified to function optimally for data where some attributes have nominal values and others are numeric. Dr. Raghavan has shown that, for the ranking method based on the Perceptron convergence algorithm [Wong88, Boll87], such modifications can be achieved. We expect

to derive similar results for other algorithms too. Following that we will demonstrate that an attribute that is non-alphanumeric can always be mapped, by clustering of similar values into groups, to a nominal valued attribute. For instance, in the case of image data, for the attribute whose domain of values consists of spatial orientation graphs, a measure that defines similarity between two spatial orientation graphs [Gudi95] can be used to place spatially similar images into the same class. Then, a new nominal attribute whose values correspond to the various class labels will replace the original attribute. Together, the above two steps would enable us to handle objects with a number of different types of attributes.

Our approach for the problem of accommodating multi-level preference structure will employ theoretical findings reported in [Boll87, Boll96]. Briefly, the strategy requires that a user, at the time of providing feedback, considers a pair of objects at a time (rather than judging each object separately as relevant or non-relevant) and indicates which of them is preferred. Using the feedback information for a number of such pairs and the feature vectors associated with those objects, suitable algorithms to derive an optimal query will be provided. The strategy, where user feedback is given by considering two images at a time and stating preference, is used in [Gudi93]. The approach currently adopted displays selected images four at a time and the user can indicate preference relative to any subset of the set of all possible pairs. Following this, the user has the option of initiating the ranking algorithm, or asking for additional images to be shown, in which case the user is wishing to continue to provide additional preferences.

The development of similar interfaces for the Web users involves a fairly complex sequence of data communication via forms, between database servers and clients. Access to image retrieval prototypes of this search facility is available from the Dr. Raghavan's home URL <http://www.us1.edu/~raghavan/>.

The approach for understanding the cost tradeoffs among different strategies for distributing ranking tasks between client and server processes will consist of developing cost models. In developing cost models, we will assume that the ranking strategies operate within the constraints of client/server architectures based on commonly adopted standards (communication protocols) such as HTTP and Z39.5. In WAIS [Kahl89], for example, it is possible to establish a "dynamic folder" that has associated with it a question or a "charter." If a server has the responsibility to respond, with a rank ordered output, to the charter of a dynamic folder, then all the information in the folder may have to be made available to it. On the other hand, if the responsibility for customized ranking resides with the client, then client must have access to information regarding the distributional characteristics of various attributes, which resides in the database at the server node. The cost models will be developed by taking such matters into account. Simulation studies will be performed using the cost model associated with each ranking strategy in order to identify an optimal distribution strategy.

3.2 Multimedia Indexing and Personal Construct Theory

Semantic attributes play a central role in Retrieval by Semantic Attributes (RSA). Semantic attributes are those attributes the specification of which necessarily involves some

subjectivity, imprecision, and/or uncertainty. We have used Personal Construct Theory (PCT) [Kell55, Kell69] as an indexing tool for systematically deriving attributes to support RSA in multimedia retrieval applications.

PCT is used to identify attributes and attribute values which can be used as key features to differentiate one type of objects from another. This information can in turn be used to automate to some extent the recognition of objects in images, one kind of multimedia objects. PCT was originally proposed by George Kelly as a clinical psychology interviewing method [Kell55]. It has since been used as a knowledge elicitation tool in building expert systems [Boos85] and to elicit expert's subconscious knowledge for image retrieval [Gudi93]. This theory is viewed as a formal model of organization of human cognitive processes. Both animate and inanimate objects with which a person interacts in everyday life constitute that person's environment. According to PCT, the objects comprising a person's environment profoundly influence his decision making process. These objects are referred to as entities or elements. A property of an element that influences a person's decision making process is known as a construct or a cognitive dimension. In other words, PCT assumes that people typically use these cognitive dimensions in evaluating their experiences for decision making. An element may possess many constructs.

The process of assigning a value for a construct on a ordinal scale to reflect the degree to which that construct is present in an element is known as rating the element on that construct. Usually, a value of one is assigned if the construct is certainly present in the element and a value of three is assigned if the construct is certainly absent in the element. To indicate a subjective neutral position, a value of two is used. However, the granularity of the rating scale can vary. A matrix that shows the element and the corresponding construct values is referred to as repertory grid. The rows are labeled with the construct names and the element names form the column labels.

In the context of image database, the following interpretations are given to the constructs and the repertory grid. Constructs are viewed as cognitive dimensions of the image domain that are useful in making relevant distinctions among the database images to facilitate retrieval. Repertory grid generation is viewed as a complex sorting test in which the images are rated with respect to a set of construct. We use the term semantic attribute to refer to a construct.

PCT experiment is carried out in two stages. During the first stage, a set of semantic attributes is discovered in the image database. The procedure for the first stage is as follows: Three randomly selected images from the image database are displayed in three quadrants of a computer display screen. The domain expert is asked to name the poles of a bipolar construct(s) vis-a-vis a semantic attribute by which images in the first and second quadrants are similar and maximally different from the image in the third quadrant. For the same set of images, the other two combinations are also considered. Then the next set of three images are shown to the domain expert and the same procedure is repeated. This process continues until the domain expert is unable to identify any more new semantic attributes. During the second stage, repertory grid is generated. The images in the database

are shown to the domain expert in a sequence. The domain expert is asked to rate each of these images with respect to each of the semantic attributes identified in stage one.

3.3 Feature-based Image Retrieval

Even though the following discussion involves one type of multimedia objects (images), it can easily be generalized to other multimedia objects. In feature-based approach, the images are represented by their semantic contents and the comparison is made between the semantic contents of the query and the images in the image database. The semantic content is the computable and low-level features, such as color, shape, texture, object centroids and boundaries, and others. These features can be derived automatically or semi-automatically.

The image retrieval (IR) system should be automatic to give an acceptable response time. The use of low-level features in the feature-based image retrieval systems, makes the approach automatic but not necessarily efficient. The use of "real" distance in the retrieval process can be computationally very expensive. For example, in the case of retrieval by color, accounting for the effect of color correlations is time consuming. The IR system should also be generic enough to be transparent when different low-level features are used with almost no or very little changes. We propose a generic and fully-automatic feature-based Content Based IR (CBIR) using color, shape, and texture properties. For example, when color is used as an indexable property, our approach directly compares the color contents of query image with those of images in the database. Low-level image properties are used to compute the real interimage distance between images with respect to a chosen sample of images. The resulting interimage distance matrix is used to generate image feature vectors for all images. These feature vectors have considerably fewer features compared to the initial image representation and the extent to which they preserve the real interimage distances among the images, can be controlled. For a given query, the feature vector is computed very efficiently by using a precomputed training set. The estimated distance to all the images in the image database is given by the Euclidean distance between the query feature vector and those of the images. Here, the image retrieval is done using these feature vectors of much smaller size. This results in a very efficient on-line image retrieval. Furthermore, the fact that the real interimage distance is preserved (to a large extent) in the feature vectors, leads to our ability to do nearest neighbor type searches. The process of deriving feature vectors and image retrieval is independent of the low-level image properties used. The complete image storage and retrieval process is done in two phases: namely, database population and image retrieval. The database population phase generates the training set in addition to computing the image feature vectors. This training set is used in the on-line image addition and the computation of the query feature vector.

3.4 Database Mining

Knowledge discovery in databases is defined as the non-trivial extraction of implicit, previously unknown, and potentially useful patterns and relationships from data [Piat91]. The data mining problem is defined to address the need for automated tools to support knowledge discovery from such large databases. For our discussion, we may view a database as a relation with a large number of tuples whose attributes are categorized as

either condition or decision attributes. In a database mining system, several classes of queries are of great importance.

Types of Database Mining Queries

Hypothesis Testing:

Hypothesis testing algorithms are fundamentally distinct from the other classes of algorithms since they do not explicitly discover patterns within the data. Instead their purpose is to receive as input a stated hypothesis and then to evaluate the hypothesis against a selected database. The given hypothesis usually represents a conjecture about the existence of a specific pattern within the database. This form of analysis is particularly useful in refining or expanding already discovered knowledge.

The hypothesis can be expressed in the form of either a logical expression in which case the hypothesis is assumed to have no antecedent; or, as a logical rule of the form "IF X THEN Y" where X and Y are logical expressions representing the antecedent and consequent, respectively. The logical expressions which makeup these two forms are defined in terms of attributes from the selected database.

The system evaluates a given hypothesis based upon the level of support and confidence it receives from the selected database. The support and confidence measures are both defined in terms of the relevant tuples contained within the database. A tuple is considered relevant if it satisfies the antecedent of the hypothesis. Thus, for a given hypothesis the level of support is measured as the percentage of tuples in the database that is relevant; and, the level of confidence is measured as the percentage of relevant tuples that also satisfy the consequent of the hypothesis.

An important issue that arises during the evaluation process is what constitutes sufficient support and confidence. The solution to this issue depends on several factors including the given user, the purpose of the request and the given data. As a result, it is only necessary to simply display the results and let the user draw his or her own conclusions.

Classification Query Algorithms:

This kind of query involves inducing a classification function (or, a classifier in terms of values of condition attributes) that partitions a given set of tuples into meaningful disjoint subclasses as defined by a user or the values of some "decision" attributes. Classification algorithms discover patterns that distinguish tuples belonging to one concept from those belonging to other concepts [Deog96]. Classification algorithms are used in two ways.

i. Classification by decision variable assumes that the concepts are derived based on the current instances of a single attribute. The selected attribute is referred to as the decision variable; and, its instances are either totally or partially partitioned into subsets, where each subset consists of instances that are identical with respect to the values of the decision variable. The constructed subsets represent the set of user defined concepts and are individually assigned a concept name. Classification algorithm is capable of analyzing the stored tuples to derive the membership conditions defining each of the concepts.

ii. Classification by example assumes that the concepts are defined in terms of two distinct sets of tuples. One set containing tuples representing positive examples and another set containing tuples representing negative examples. Specifically, the user, through a sequence of SQL queries, specifies the tuples representing positive examples and the tuples representing negative examples. The labeled tuples are then analyzed by the classification algorithm to determine the membership conditions defining the two concepts.

Characterization Query Algorithm:

Unlike a classification query, a characterization query describes common features of a class regardless of the characteristics of other classes.

Characterization algorithms discover patterns which characterize the tuples belonging to a single predefined concept. Like the classification algorithms, the characterization algorithms also analyze tuples based on their membership to a specific concept. However, the sets of tuples analyzed by the two classes of algorithms are different. As noted in the previous section, the classification algorithms compare tuples from distinct concepts. However, the characterization algorithms compare only tuples of a single concept. The implication is that the classification algorithms may not discover all the commonalities among tuples of a single concept; and, the characterization algorithms may discover commonalities which are not unique to a specific concept [Cai91].

This algorithm also allows the user to specify concepts in terms of either a decision variable or a sequence of SQL queries. In the case of a decision variable, a set of concepts are characterized and in the case of the the SQL queries a single concept is characterized.

Association Query Algorithm:

Data dependencies are useful to determine associations among values of certain "condition" attributes. These algorithms discover patterns that are more general than those discovered with either the classification or characterization algorithms. The analysis itself, unlike that performed by the previous algorithms, is not based upon a set of user defined concepts.

The association algorithm discovers associations among the attributes of the given database. Associations of this type are said to exist when the same attribute values occur in multiple tuples. This form of association is likely to occur frequently within a given database. However, many of these associations will have relatively little support given the current state of the database. To eliminate them from consideration, and thus allow the algorithm to operate more efficiently, the user is required to specify a minimum support requirement.

The level of support for an association is defined as the percentage of tuples which contain an instance of the association. It is important to note that the enforcement of a minimum support requirement does not require the algorithm to determine the actual support level for every association existing within the database. This fact is the result of the following property: if a set of K attributes does not satisfy the support requirement, then any superset

of the set K will also not satisfy the requirement [Agra96]. As a result, the algorithm for discovering associations is implemented in terms of a bottom-up, iterative procedure.

Clustering Query:

We call unsupervised partitioning of tuples of a relational table a clustering query. Clustering queries may be helpful when labeling of a large set of tuples is deemed too costly and time consuming. Instead, a classifier may be designed on a small, labeled set of samples, and then tuned up. The task of clustering is predicated on the assumption that given any two tuples a measure of distance can be computed.

3.4.1 Rough Sets

Rough set theory was introduced about a decade ago by Z. Pawlak. The rough set methodology is highly promising for database mining in many business and scientific domains. A review of the literature shows that the hypothesis testing and classification queries can easily be solved by the rough set methodology. However, the following problems are not yet adequately addressed and will form the focus of our research.

Incremental rough approximation:

This is a must feature if the background knowledge is dynamic. We will develop evolving rough classifiers by devising a relational representation of classifiers with a composite field containing frequency counts needed to measure their worth.

Closeness of two rules:

Determining the nearest rule, in the case that the description of a given object does not match one of the learned object classes, is a key factor in enhancing the performance of a rough classifier. By introducing similarity measures we will enable rough set methodology to be used for clustering queries.

Characterization query:

Current application of the rough set methodology to characterization queries suffers inability to use explicit structure, such as a hierarchy of concept dependencies (e.g. "sedan" is an "automobile"). We will extend existing results in the use of concept hierarchies to a rough set-based approach for characterization queries.

3.4.2 WebMine

Database mining applications require the knowledge discovery capabilities to be provided within client/server environments. Until recently, remote access of data often required dedicated computer networks and customized display stations. In the context of the Internet, the trend is towards having software modules for accessing databases built upon the Common Gateway Interface (CGI) protocol and the HTTP. Users can access and retrieve information maintained at servers by using browsers such as the Netscape Navigator or the NCSA Mosaic at the client side of the communication channel.

We have implemented a prototype, called *WebMine*, to provide data mining tools. *WebMine* is a system consisting of clients, servers, CGI programs and a DBMS. The client is the machine from where the user is accessing the *WebMine* system. It plays two roles. First, it lets the user send inputs to the different CGI programs. Second, it displays HTML documents received from the server. The server is the machine on which the CGI programs are stored as well as the HTML documents. The server is responsible for invoking the CGI programs and sending the HTML documents to the client browser. In general, CGI programs parse user input and execute specific algorithms in response to the user's input. That is, they can create either an HTML document or an HTML form as output. The executed modules correspond to the supported data mining algorithms. The execution of these algorithms requires a CGI program to formulate a series of database queries. The formulated queries are evaluated by the underlying DBMS. At present, hypothesis testing and association query algorithms are complete. Rough set based algorithms for classification queries and characterization queries should be complete in a few weeks.

The widely used Java programming language could be used in this project to develop algorithms for our query classes. These modules will be architecture-neutral, so that they can be portable for multiple platforms in heterogeneous, distributed networks. The object-oriented approach of Java is well matched to the distributed client-server tool requirements of this project. Using Java, applets, which are embedded in an HTML document itself, can be developed. Besides providing visualization and processing utilities, applets can be used to offload much of the computation related to mining tasks to the client machine. We will also investigate the incorporation of data security measures as applets. Furthermore, the multithreading capability of Java will allow concurrent activities be handled; activities that might include progressive display and visualization of portions of a data set while the data are being downloaded.

3.5 Application Partitioning for Distributed Computation

When confronted with running a computationally intensive task, today's user turns to a cluster of workstations rather than a massively parallel processor (MPP). Workstation clusters are highly available, fault tolerant, and easy to maintain. In this section, we address efficient methods for application partitioning for distributed computation with special emphasis in information retrieval, data mining, and image processing to complement the corresponding research efforts in this proposal. Our ultimate goal is to develop a library of procedures optimized for the workstation cluster environment for on-line applications described in sections 3.1 to 3.4 and Section 4 of this proposal.

The biggest challenge in utilizing a workstation cluster as a high-performance computer is the inter-process communication. Efficient methods are needed to parallelize large computations that minimize the communication and maximize the execution concurrency. The workstations may have different speeds, so the partitioning of a large task must account for heterogeneity. There is no single approach for partitioning of computations which is best for all types of tasks. In general, application programs may possess arbitrary control and

data dependency structures among their subtasks, and processes may communicate and synchronize with each other. While this allows for the most general form of parallel computation, in practice, we must be more focused to identify the class of parallel computation structures suitable for workstation clusters.

Two typical computation structures are the Single Program Multiple Data (SPMD) model and the functional parallelism. In the functional parallelism, different subtasks of a computation may have different algorithms. In the SPMD model data parallelism is exploited by partitioning the data space between several processors all of which execute the same code. This type of parallelism arises often in the proposed areas of application, and due to their widespread applicability, they have been extensively studied in the literature [Chen92,Karp87,Li91]. The major advantage here is that only one copy of a sequential code needs to be written which is then executed on all of the processors.

Even when we focus on SPMD computations, we encounter a wide spectrum of possible approaches to partitioning the computations. For example, if the task does not require synchronization or communication between subtasks, then the problem is relatively simple since whenever a workstation is free we may schedule the next subtask for execution. A more general case is when the subtasks of a computation need synchronization between themselves in order to exchange data. This arises, for instance, in a lot of computations involving the solutions of differential equations, low-level image processing functions, or database management systems. In such computations, all subtasks start execution at the same time. Once the execution has started, different subtasks may execute at different speeds since different workstations may have different processing rates. When a subtask reaches its synchronization point, it waits for the other subtasks to reach their synchronization points. When all subtasks reach their synchronization points, they exchange data for the boundary points, and then again they start the next iteration at the same time. As a result, the subtask that executes on the slowest workstation requires the longest execution time. The task is not completed until the subtask with the longest execution time is completed.

For this second type of tasks, we developed an optimal scheduling algorithm for determining, simultaneously, the best subset of workstations to be used, and the best way to partition the task between the available workstations [Efe95]. This result is theoretical, in the sense that it considers a general task system without a detailed consideration of the algorithm of the computation. In practice, when we closely look at the algorithm being parallelized we may be able to detect special reasons for false bottlenecks and identify opportunities for improving parallelism.

To analyze such properties of a computation, we propose to extend the above algorithm to partition the computations in a way to minimize data dependencies between subtasks. There are two types of partitioning that can be performed on a parallel program: partitioning based on control parallelism or partitioning based on data parallelism. Control parallelism is obtained by partitioning the program into parallel tasks, each with different functionality, so as to balance the amount of workload assigned to each processor. Data parallelism is

achieved by partitioning the data space between processes, and may be used in conjunction with control parallelism in order to further reduce communication overhead and increase data locality.

In this research we focus on the SPMD model of computation where the algorithm is the same for all the nodes in the network, but different sections of the data space is partitioned between workstations. Accordingly, the main emphasis of research will be on exploiting data parallelism. We have several reasons for this focus:

1. In MIMD parallelism, control synchronization is needed in order to enforce the correct order of execution between subtasks of a program running in parallel, so as to guarantee the consistency of the parallel execution. If implemented in workstation clusters, such control signals can lead to significant delays. On the other hand, SPMD model of computation does not require control synchronization between subtasks except at the beginning or at the end of the computation (synchronization due to data dependency may still be needed as mentioned above).
2. In MIMD parallelism, or control parallelism, schedule optimization assumes that the running times of different modules in a program are known. While several strategies have been suggested to estimate the running times of program modules, none of them guarantees the accuracy of estimations. To illustrate the difficulty of estimating the running time, consider the following extreme case:

```

int x; read(x)
Repeat
if x is even
    then x=x/2
    else x=3x+1;
until (x=1)

```

There is no way the running time of this algorithm can be estimated by inspecting the value of the input parameter. One can easily check that if initially $x=120$, or $x=1048576$, then the algorithm will repeat for 20 iterations. For $x=128$ it will terminate in seven iterations. It is a well known conjecture that this algorithm will terminate for every initial value of $x>0$.

In SPMD computations, it is data space, and not the algorithm of the system, that is partitioned between workstations [Gupt92]. Thus, running times of different processes can be made equal by partitioning the data space equally between workstations. There are several computations (e.g. "embarrassingly parallel" computations) for which this problem of partitioning the data space is fairly straightforward. There are other computations for which this problem is more challenging, and still others for which the problem is quite difficult or impossible (the algorithm of Figure 3 appears to be an example in the "impossible" class). Never the less, partitioning the data space alone is much easier than partitioning the algorithm as well as the data space.

3. SPMD model of computation appears to have a wide range of applications. For example, most of the SIMD algorithms developed for massively parallel computers (which is a rich class of algorithms) can be converted to SPMD algorithms by simply letting each workstation to emulate a large number of processing elements. Then, schedule optimization for a workstation cluster will involve supplying the right input data to the

right workstation at the right time. This can be optimized off-line due to the deterministic nature of SIMD computations [Koc92,Efe95b].

4. Our final reason for considering SPMD computations is that the algorithms in our focus area lend themselves for efficient implementation in the SPMD model. These algorithms have such widespread applications that parallelization of this class of computations on workstation networks should alone provide enough reason to pay more attention to workstation clusters as parallel architectures.

In order to apply this type of approach to a workstation cluster, the main focus of research needs to concentrate on data itself, although of course knowledge about the algorithm of computation is also needed at least to decide which data is needed, where, and when. Accordingly, the following are some of the major components of the proposed research in parallelizing an SPMD computation.

3.5.1 Data Replication

In many cases data replication may reduce the amount of communication overheads, although it increases the amount of memory requirement for the processors. A datum is called input datum if it is only read during the entire computation. Instead of having only one copy of the input data, replicating it to all processors which read it reduces the communication overhead. This is because this data is to be read only and it is much less expensive to replicate by efficient communication primitives (e.g., broadcasting) than to fetch it one by one from a remote processor during the run time. When only limited data replication is possible, e.g. due to limited memory size at each workstation, we conjecture that an approach based on bin-packing algorithms should prove to be useful for optimizing data replications. We will address this problem and develop an efficient method to find the optimal replications. Input parameters of the problem will necessarily be the available communication primitives in the system and their corresponding timing requirements, such as broadcast, point-to-point communication, etc.

3.5.2 Run-time data communication

During the computation of an SPMD program, communication may be needed in order to exchange boundary values between different processes so that independent programs may adjust or tune their parameters, or even decide to halt [Atal92]. Besides these, if the task dependency structure is static, then we can use a good heuristic algorithm from the literature to minimize the communication [Kris96]. If however different instances of the SPMD program branch differently, then the execution of a branch in one process may depend on the selection of a branch in another process. A typical example of this arose when parallel ray tracing was performed on a set of processors none of which had enough memory to hold the entire image [Whit92]. Here, branching of computation depends entirely on the input scene, and arbitrary communication patterns between subtasks may be generated during a computation. In this work, a set of "learning heuristics" are used which predicted the remaining patterns of communication based on the observed pattern after some time of initial computation. We will investigate the applicability of similar approaches for other types of computations.

3.5.3 Optimizing number of processors and the message size

In workstation clusters, communication could be very expensive. Often, the cost to send one datum from one processor to another could be more expensive than the cost to compute several hundreds of floating point multiplications. Using more processors to execute a task may decrease the computation time but the communication time may increase. Thus there is a tradeoff between the communication overhead and the waiting time. The optimal number of processors is defined as the number of processor by which the overall completion time is minimized. For this optimization problem, we will investigate methods such as those recently used in the PhD thesis of Dr. Efe's student [Utha96]. We express the overall execution time as a function of the number of processors used, the message size, the problem size of the algorithm and the architecture parameters. Then we find the optimal number of processors by taking derivative of the cost function with respect to the communication parameter.

The image processing applications of this research will support the missions of the RVC. Many applications in environmental research, atmospheric sciences, agriculture, oil/gas production, etc require processed satellite images rather than the raw data. The processing required typically involves such tasks as feature extraction, change detection, image classification, image registration, texture analysis, image enhancement, filtering, sub-pixel extrapolation, and wavelet and multiresolution analysis. Due to the massive amounts of data involved in fusion of satellite sensor data with in situ sensor data, serial processing is not feasible for many types of fused images.

Opportunities for parallel execution of image processing/understanding tasks are detailed in a number of papers (see for example [Wall92,Sieg92]). Typically, these applications use low level functions such as pattern matching, linear or non-linear window operations for transformation of an image into another image, and other numeric and symbolic computations for high level image understanding tasks. The pattern matching task is inherently parallel as each pattern being matched against an image sequence can be processed by a different workstation. There are other tasks, such as those based on histogramming, that require some amount of communication between the processes. Also, numeric and window based operations typically require communication between processes. While efficient algorithms have been developed for many of these tasks for massively parallel computers, the relatively small communication bandwidth available for workstation clusters implies a different level of task granularity for optimal performance, and these earlier algorithms do not necessarily have a straightforward translation for workstation clusters. The proposed research will require a careful analysis of alternative ways for partitioning the data space between workstation.

Much of the work done in information retrieval and data mining use certain low level primitives in searching, query processing and pattern matching, and these tasks are inherently parallel. Additionally, tasks involving image data bases often use such computations as image classification, texture analysis, etc, are similar to those in the

previous paragraph and thus the corresponding parallel algorithms can be used in both application areas.

Additionally, information retrieval often involves search in multiple databases and offers interquery parallelism. However, such opportunity brings with it challenges in optimizing queries running on multidatabase systems, an underlying notion of which is that different components in the multidatabase have no notion of cooperation [Ozsu91]. Each database may have its own transaction processing services (transaction manager, scheduler, recovery manager). As a result, local cost functions are not known a priori, and they cannot be communicated to the site of the client. This makes it harder to carry out query optimizations, but intelligent agents can be used to estimate these cost functions based on historical data. This information can be used for query optimization not only for execution on a cluster of workstations, but also for execution across the network.

4. Regional Validation Center Enhancement

4.1 Background

USL has been designated by the NASA Goddard Space Flight Center as one of four sites in the United States for a Regional Validation Center (RVC). The RVC receives transmissions from satellites which cover the Southern United States and the Gulf of Mexico region. Using a state of the art earth station with advanced computer hardware and software, it directly collects and processes the massive amount of data transmitted by current and future satellites in real time. The processed data can then be widely disseminated to promote development of new applications of satellite data by government agencies, university research scientists, and industry. The availability of processed satellite data is expected to have a major impact on research and economic development, especially those that pertain to energy and environmental applications, in the region. Data collected and analyzed by the RVC will be used by the Information Center as its local repository.

4.2 RVC Concept

The goal of the prototype RVCs is to foster the self-supporting use of environmental and earth resource data (from satellites and other sources) by regional institutions including state and local government agencies, universities, consortia, and commercial companies. The objectives of the prototype RVC effort are to:

- Establish several prototype RVCs at existing institutions representing a diverse range of applications,
- Quickly provide the RVCs with flexible, scalable information systems which incorporate unique NASA-developed beta-test information technologies for cost-effectively making large and complex remotely sensed data sets useful and accessible,
- Promote the establishment of self-sustainable public and private sector working relationships and programs that usefully apply NASA satellite data at the local community level,
- Refine and transfer NASA information technology through collaborative testbedding,
- Use RVC-created in-situ and ancillary data bases to support the calibration and validation of NASA satellite data, and
- Incorporate the RVCs' applied research results into sharable global environmental knowledge bases.

4.3 RVC Initial System Functionality

In order to accomplish these objectives NASA has installed some equipment, on loan, to USL for an initial period of 5 years. This equipment includes a parabolic antenna and antenna controller, a Pentium PC computer to act as an ingest system, a Unix-based workstation, and a mass storage device. The equipment was installed in the Fall of 1996 and is currently operating using the initial, beta release of the software. We anticipate that the

system will go into full production mode with the next release of the software in the summer of 1997.

4.3.1 Ingest, catalog, and store satellite data

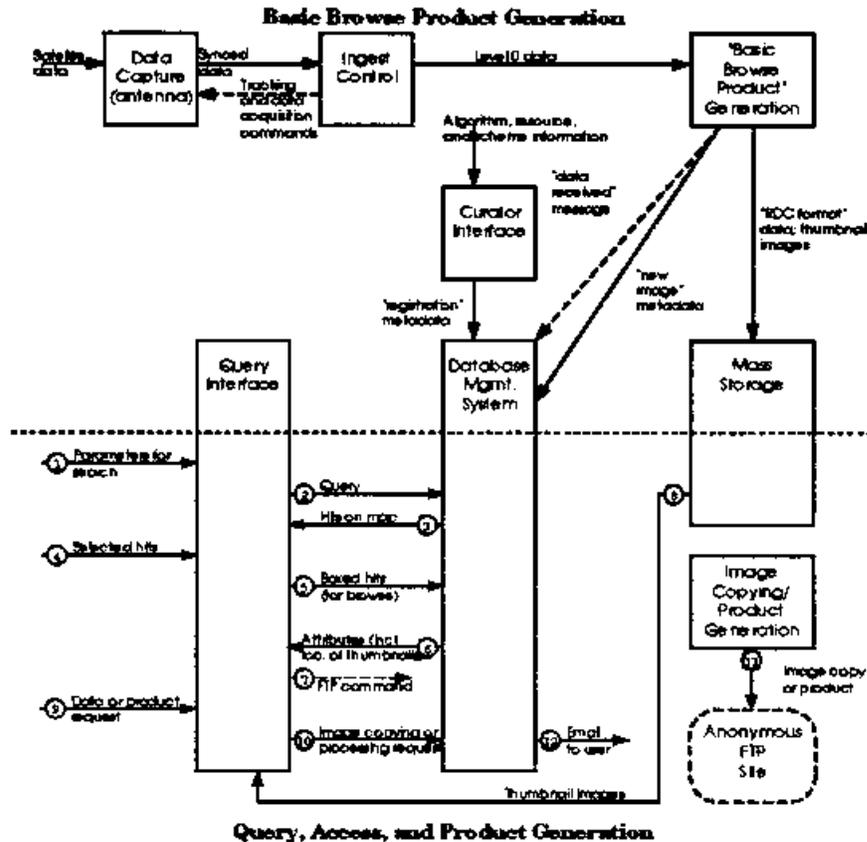


Figure 1 - Data Ingest, Cataloging, and Storage

The ingest system will maintain information on satellite locations and data downlink schedules, and control the tracking of the antenna during passes. The data stream will be stored in its raw format on the ingest system. Once a data stream has been captured the data file will be transferred to the Unix-based workstation for conversion into "RVC format". This process creates separate data files corresponding to the sensors on the satellite. The resultant data is then stored in the mass storage system along with reduced resolution "thumbnail" images. As the input data stream is ingested, metadata (e.g., platform/sensor/channel, time of acquisition, "basic browse products"¹) are computed or extracted and placed in the object oriented database management system. (See Figure 1)

¹ Basic browse products include any products (the result of running certain algorithms on a given data set) that need to be generated for each and every data set that is received.

4.3.2 Support registration of algorithms, schemes, and resources

The RVC provides an interface through which a system curator enters into the database management system the names of algorithms and the information needed to run the algorithms. This includes the types of images to which the algorithms are relevant. The curator interface also allows the system curator to register a "scheme" (e.g., "water", "forest", "urban") in the database management system. Finally, the curator interface allows the system curator to identify the computing resources associated with various system processes such as available CPU resources and anonymous FTP servers.

4.3.3 Support database queries through a graphical user interface

NASA will provide each RVC with a graphical interface through which users identify search parameters for data cataloged by the database management system. The interface provides for search by temporal, spatial (location and/or resolution), spectral (bandwidth), platform/sensor/channel, and content-based parameters. Each time the query tool is invoked, it queries the database to insure that the interface always reflects the most recent information in the database. (See Figure 1)

The results of each query are provided in the form of a map showing the location of database "hits"; available data sets which fulfill the constraints of the query. For each hit, an image identifier is shown as well as optionally displaying a "thumbnail" of each image. Once an image has been identified through a database query, the user is allowed to request any of the applicable products for this image. The requested products are generated and copied to an anonymous FTP area. The user is informed via electronic mail as to the location of the product.

The query tool also provides a display of all algorithms registered in the database which are applicable to each image via a pull-down menu. If content extraction algorithms have been run and schemes have been produced, an additional menu is provided for content-based results. Throughout a session the query tool maintains a query history to aid in storing and re-executing complex queries.

4.3.4 Integrate planning and scheduling capabilities

The RVC provides a planner/scheduler/dispatcher which controls disk space management for data ingest, basic browse product generation, and menu-driven product generation. The planner queries the database to determine what resources are available and determines the resource allocation to be used to accomplish the goals. (See Figure 2)

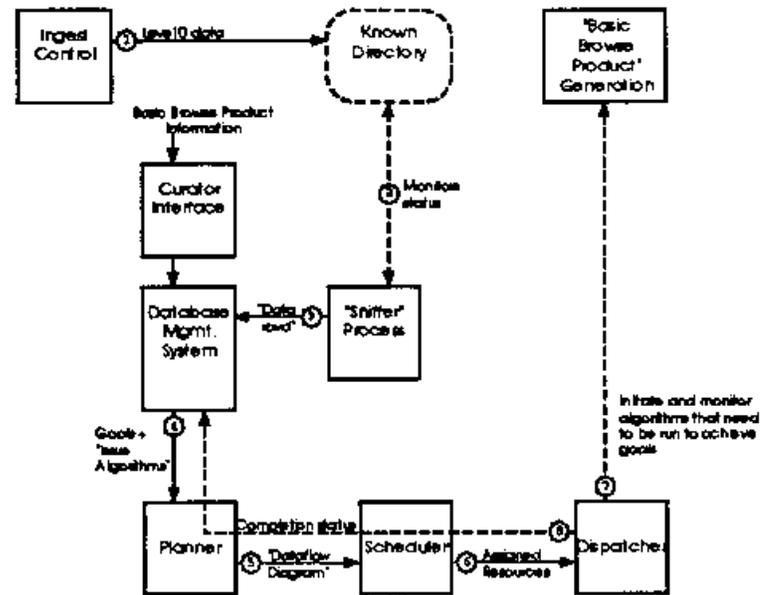


Figure 2 - Planning / Scheduling of Data Ingest

4.4 Rationale for Enhancement

The purpose of this proposal is to enhance and extend the capabilities of the RVC thereby making it available to more researchers, students, and collaborators. This will be accomplished by building upon the NASA supplied hardware and software and taking advantage of the scalability of the RVC design. A robust user interface which is easily accessible through the World Wide Web will also be developed. This will allow users with a wide variety of concerns ranging from scientific research to commercial applications and public education to access the RVC.

In the current RVC configuration the single Unix workstation supplied by the NASA Goddard Space Flight Center is responsible for running the object-oriented database management system, the planner/scheduler processes, managing the mass storage system, supporting interactive users of the graphical query tool, supporting interactive users wishing to browse the data sets available in the RVC repository, acting as the anonymous FTP repository for the RVC, and serving as the CPU server for all basic browse product generation and advanced algorithm processing. Without this enhancement the RVC will be constrained by the capabilities of this one machine and will have a more limited impact. With this enhancement the potential applications of the RVC expand from a data collection and dissemination platform to an attractive platform on which further research programs can be built.

The National Wetlands Research Center (NWRC), which is housed in USL's research park, is a partner of the University in the NASA/USL RVC project. The mission of the NWRC is to provide national leadership in biological research and development related to protecting, restoring, and managing natural resources, with an emphasis on fish, wildlife, and wetlands

in the South. Currently the NWRC focuses on wetland, forest, and animal ecology; spatial analysis; and information and technology transfer. The NWRC has a staff of approximately 150 scientists and support personnel. The NWRC will be instrumental to the success of the RVC in two areas.

First, the NWRC has an extensive client base to which the services and products of the RVC will be very valuable. By partnering with the NWRC the university will be able to take advantage of this strength of the NWRC to increase both the visibility of the RVC and the dissemination of RVC data and products. Second, the NWRC is constantly collecting in-situ data which will be combined with the data collected by the RVC. The in-situ data can be used to perform ground truthing on the RVC data and can also be used as additional input to product generation algorithms.

In order to ensure that the NWRC and the NASA/USL RVC can interchange data effectively a high-speed network link is included in this proposal. The cost of the link will be shared with the NWRC. The Center's director, Dr. Robert Stewart, has shown great interest in the NASA/USL RVC project. He has written a letter in support of the RVC and this proposal which is included in Appendix .

USL is affiliated with the Louisiana Universities Marine Consortium (LUMCON), a statewide organization involved in a variety of marine research and educational programs. Louisiana's coastal offshore waters, coastal habitats, and valuable living resources form the focus of LUMCON's research emphases. Current research programs integrate the following themes: environmental processes of coastal change, biological productivity, production and development of Louisiana's fisheries, environmental effects of energy-related industries, and interactions of the Mississippi River with the Gulf of Mexico. Senior scientists at LUMCON hold adjunct appointments to the USL Biology Department. LUMCON's executive director, Dr. Michael Dagg, has also written a letter of support which is included in the Appendix .

The researchers at NWRC and LUMCON are in a unique position to provide the RVC with in-situ observations to validate and enhance the interpretation of remotely-sensed data such as satellite images. In addition, current and future research programs at those institutions could benefit greatly from the incorporation of satellite data. Many offshore oil installations are also equipped with atmospheric sensors which could be integrated into the RVC's databases to enhance the interpretation of the satellite data and the accuracy of the researcher's models.

Other professionals with ties to the offshore oil industry who have been briefed on the capabilities of the NASA/USL RVC have also shown great interest. Several letters of support from industry representatives have been included in the Appendix .

5. Schedule of Work

In year one, the development of the Information Center will proceed along the plan as outlined in Section 2.2.4. In particular, the Advisory Board will be constituted. The investigators, the information professionals, the research librarians, with input and advice of the Advisory Board will identify the client base and the prototype resources. Prototypes for the software tools will be designed and developed in year one, and refined by incorporating research results in years two and three. In year two, the metadata will be written for those data sets currently lacking such. Also in year two, the graphical user interfaces for the tools will be refined. In years two and three, the tools for supporting advanced features such as interoperability among various resources will be provided.

The research program will report the analysis and results on a periodical basis, as dictated by the Administrative Plan established by all projects in the EETAP package. The research goals are as specified in each of the sections (Sections 3.1-3.5). The research results will also be incorporated into the software tools (Sections 3.1-3.5). Results from Section 3.5 will also be incorporated into the RVC planner/scheduler activities and tested as they become ready, during years two and three.

The RVC enhancement of equipment will be performed in years one and two, to support the work of the other components. In years two and three, the analysis applications software from Section 3.5 will be incorporated into the RVC planner/scheduler and tested.