

<b>ABSTRACT</b> .....	1
<b>1. PROJECT DESCRIPTION</b> .....	1
1.1 Research Program - Executive Summary .....	1
1.2 Project Outline .....	6
1.3 Proposed Scenarios .....	7
1.4 Content-based Retrieval for Model Generation and Validation .....	14
<i>1.4.1 Enabling Technology</i> .....	14
<i>1.4.2 Proposed Solution</i> .....	16
<i>1.4.3 An Example</i> .....	17
<i>1.5 System Architecture</i> .....	19
1.6 Innovative approaches for data production. ....	21
1.7 Support for the Earth System Science Users .....	22
1.8 Interoperability .....	23
<b>2. PARTICIPATION IN THE WP FEDERATION</b> .....	24
2.1 Federation Objectives .....	24
2.2 Participation Methodologies .....	24
2.3 Expected Contributions .....	25
<b>3. METRICS</b> .....	25
<b>3.2 Earth System Science Metrics</b> .....	25
<b>4. MANAGEMENT APPROACH</b> .....	26
<b>5. PERSONNEL</b> .....	27
<b>6. PROPOSED COSTS</b> .....	29
<b>7. COOPERATIVE AGREEMENT PAYMENT SCHEDULE</b> .....	30

## **ABSTRACT**

We propose developing a prototype system to explore the effects of environmental factors on public health through the retrieval and analysis of remotely sensed images. The methods of content-based retrieval are applied to spatial and temporal data in order to investigate factors that affect public health risk such as temperature, precipitation, vegetation, and land use patterns. The information obtained from remotely sensed images is then used to validate the risk assessment models for infectious and environmental diseases or hazards such as Hantavirus Pulmonary Syndrome, air pollution, fire ant infestations and malaria. The prototype system is capable of participating in and communicating with a federation of ESIP servers. The objectives of the proposed system are to: (1) establish a new paradigm for investigating risk factors in the public health community which integrates spatial data mining and statistical modeling into a novel content-based search and retrieval technique on remotely-sensed data, (2) generate and provide risk assessment maps for public health problems and diseases with much finer temporal and spatial precision and coverage than has been previously possible, (3) interact with other ESIP systems to share and discover data that is related to the development and validation of public health risk assessment models.

## **1. PROJECT DESCRIPTION**

### **1.1 Research Program - Executive Summary**

Interdisciplinary research provides great opportunity for combining information and knowledge from multiple disciplines. An excellent example, of relevance to the WP federation, is the emerging field of environmental epidemiology, due to its use of satellite imagery, derived products, and GIS data. Researchers seek to identify environmental factors that contribute to the spread or increased risk of disease and that, in general, affect public health. By integrating data from earth observation imagery that describes weather patterns, topography, land use, vegetation, etc., with public health information, environmental epidemiologists seek to generate models predicting the outbreak and spread of diseases such as Hantavirus Pulmonary Syndrome, Dengue fever and malaria.

This interdisciplinary research requires new tools and methods for searching, retrieving, manipulating and integrating data from multiple sources to generate, validate and apply new public health models. Furthermore, interoperability within a federation of distributed data servers and clients is critical.

The goal of this proposal is to support generation, validation, and large-scale application of earth sciences models in a federated environment. We propose an open system that combines techniques for spatial data mining, statistical analysis, and content-based search. Although we initially focus on applications within environmental epidemiology, we will develop a domain-independent infrastructure which is applicable to other interdisciplinary research areas.

The proposed project will explore techniques to:

- Locate available satellite images, geo-spatial data and other data types from a set of federated image archives,

- Select sub-regions for analysis using content-based retrieval techniques,
- Determine the relative importance of various environmental factors for a specific dependent variable through data mining, knowledge discovery, and interactive query refinement,
- Scale the modeling, spatial data mining, and content-based retrieval algorithms to a large amount of data (in excess of 280 GB/day),
- Develop a flexible system that is easily adapted to a wide variety of models by different users.

We illustrate the proposed technologies in the following example. Consider the following scenario for a rodent-borne viral disease such as Hantavirus Pulmonary Syndrome (HPS; see below for details) outbreaks of which we would like to predict. The steps in modeling HPS epidemiology are as follows:

1. The scientist suspects that the location of a house (proximity to a wet grassland makes it more prone to large populations of mice) and recent temperature and moisture patterns are the most important factors for predicting the outbreak of HPS. The user formulates a content-based query by composing a search consisting of (1) the co-occurrence of a specific texture and spectral pattern (to locate wetlands and houses), (2) a weather pattern specified by a time series, and (3) ground moisture and temperature levels. The results are ranked based on the similarity to each of the three criteria using user-assigned weights. This query can be viewed as a precursor of a model that can be used to predict outbreaks of the virus. The user compares the results with the historical database, which contains the geographical coordinates of each disease outbreak.
2. The user would like to test additional factors that may contribute to the outbreak of hantavirus. The user formulates a new query by adding vegetation index, ground moisture level, and elevation, slope and aspect to the model. Through a process of iterative refinement which compares the query results with the historical database, the model is revised.
3. Although houses are located with high resolution images (5 meters or below), these images may be prohibitively expensive or simply not available. Furthermore, the spatial coverage of precipitation and temperature data is not complete. Consequently, alternative methods need to be developed to substitute the missing data. For example, lower resolution data, such as Landsat TM data (with 30 meter resolution), can be used to infer the location of houses by extrapolating from neighboring areas. The rainfall and moisture data can be extracted from the vegetation index and water vapor channel provided by many satellites. The system will thus determine which set of data to use for a given location based on a set of substitution rules and budget constraints. In general, the data will reside at multiple sites which are managed by different archive centers.
4. After the model is generated, the user may consider a wider spatial area and temporal period. Wide-scale application of the model presents challenges in scalability of the model and the corresponding search, retrieval and data

processing algorithms. In particular, this introduces new concerns regarding the possible unavailability of data for wide-scale temporal and spatial application.

The proposed technology makes the operations described above readily accessible and generalizable to a variety of domains including as agriculture, forestry, fishing, and transportation. In the first two steps above, we see the critical role that content-based retrieval can play in effectively utilizing earth observation imagery. In the third step, we see that interoperability within the federation of data archives is essential. The last step, highlights the importance of scalability in applying models. Furthermore, we address the problems of missing data through the automatic application of data substitution methods which are provided in the system.

The discussion below details the proposed project in terms of the new and innovative capabilities to be developed, including modeling facilities, data access methods, and new services to be provided to the earth sciences community.

## **New Capabilities**

### **(1) Model Generation, Validation and Proliferation**

Model generation is the process of identifying relations between a set of independent variables (predictors or covariates) and a set of dependent variables (responses). Presently, there are several methods for generating and validating models using spatial data. These methods can be grouped into three main categories: 1) statistical analysis, 2) spatial data mining and 3) content-based search. We briefly describe state-of-the-art for each.

1) Numerous packages supporting standard statistical techniques are available. They include statistical systems such as SAS, SPSS, BMDP, and statistical programming languages such as S, MATLAB, ENVI/IDL and MAPLE.

2) Spatial data mining is an emerging discipline which extends traditional databases data mining operations (such as discovery of association rules, attribute focusing, clustering) to Geographical Information Systems (GIS) and other spatial data. The objective of spatial data mining is to provide methods for efficiently analyzing large quantities of spatial data.

3) Content-based search is a recent technique for extracting information from nontraditional databases (such as image, video and multimedia archives). Examples include the IBM QBIC project (Flickner et al., 1995) and the Virage system (one of the Informix datablades) (Bach et al., 1996), which provide methods for image retrieval based on features such as texture, color histogram, and shape. The UCSB Alexandria project (Ma & Manjunath, 1996) extends the functionality to the retrieval of images based on local texture features. The SaFe (Smith & Chang, 1997) and VisualSeek (Smith & Chang, 1996) projects from Columbia University and the Blobworld (Carson et al., 1997) and Bodyplan (Forsyth & Fleck, 1997) projects from UC Berkeley support image search based on spatial configurations of objects and regions.

None of the above approaches is individually sufficient to generate and validate earth science models. Statistical tools lack the capabilities of interfacing with large GIS databases, and cannot extract features and compute covariates from remotely sensed

images and derived products. Spatial data mining methods achieve efficiency at the expense of sophistication of the resulting algorithms; current spatial data mining functions are inadequate for modeling complex phenomena. While content-based search is useful for locating and retrieving the data relevant for a particular analysis, it does not perform statistical analysis.

We propose a new methodology which provides a common platform for combining content-based retrieval, spatial and time-series data mining, and statistical modeling. This methodology uses an object-based approach to encapsulate those salient features (such as texture and contour), spatiotemporal patterns, and spatiotemporal relations that are relevant in constructing a statistical model.

This platform is used to construct, refine, and validate models to be used in predicting locations of high risk. A set of new techniques for content-based query evaluation and retrieval indexing are also proposed to ensure the scalability of the proposed methodology.

The modeling platform will be packaged as a set of readily distributable components (JavaBeans), which can be combined with existing user-interface components, allowing a user to quickly package a generated and validated model as a stand alone application.

## **(2) Progressive and Distributed Access across Heterogeneous Data Archives**

The key challenges to accessing data in a heterogeneous environment are to: (1) locate relevant information sources, and understand the metadata that are used to describe these information sources (2) provide a unified query interface for a large number of databases, (3) evaluate the queries in a distributed fashion, (4) merge the query results from these sources, when the evaluation criteria might not be identical.

In recent years, a number of standards have evolved for referencing and accessing data in a distributed, heterogeneous environment. These include standards for describing metadata, such as FGDC's CSDGM, standards for storing spatial data, such as HDF/EOS, standards for data transfer, such as SDTS, standards for search and retrieval, such as Z39.50, and others.

Several research projects have addressed the problem of accessing data in a heterogeneous environment, including the Garlic project at IBM Almaden Research Center and the STARTS project at Stanford University. In both systems, a metasearcher is used to dispatch queries to databases that might contain results. In the case of Garlic, a wrapper is used to hide the proprietary detail of each database. A unified protocol, Z39.50, is used to access different resources in STARTS. These standards and research projects have developed addressed interoperability on metadata, but have left unsolved issues of interoperating on data abstracts such as low level features (e.g., texture, spectral histogram) or objects.

We note that in the Earth Science arena there exists a distinct dichotomy between metadata and data: none of the standards provide intermediate abstraction layers to describe the underlying data sets. For example, an image can be described in terms of a compact sets of features (such as texture, local spectral histogram, etc.), or an even more compact set of semantic descriptions. When available, searches can be very efficiently performed on the semantic description; when content not contained in these descriptions

is required, queries can be formulated in terms of lower abstraction layers (for example, by specifying a texture feature). Higher abstraction levels correspond to more compact representations. Thus, searching at the semantic level results in reduced computational burden on the server. Also, searching at higher abstraction levels allows us to more efficiently prune the search space and produce a smaller set of query results. This reduces communication costs.

We propose developing and extending the WP Federation to include a hierarchical description of content, called InfoPyramid. InfoPyramid includes feature and semantic level representations of the data. InfoPyramid is an object-based model of content, and includes both an abstract data model and software components, which can be distributed on-the-fly. This approach allows searches to be performed at sites that do not store all levels of the InfoPyramid data representation. The InfoPyramid will be integrated with interoperability standards such as the Z39.50, CEOS CIP, EOSDIS Core System and FGDC metadata. Furthermore, the rapid retrieval and delivery of data will be facilitated by use of the Space and Frequency Graph (SFGGraph). This is a novel technology which enhances storage and access of large volumes of image data.

Of particular importance when applying a model to different geographical areas or to different time frames is the capability of identifying and substituting datasets for those originally used for development. For instance, some datasets might not be available for all regions. In this case, we propose a rule-based system that would identify alternative datasets or functions of alternative datasets, and substitute them for the missing ones. For example, rainfall data might not be available for a region of interest, but could be successfully replaced by a function of the relative greenness and of the soil moisture index derived from low-resolution satellite images such as those acquired by the NOAA-AVHRR sensor. Such data substitution will be based on a cost/benefit analysis incorporating data quality and delivery cost measures.

### **New Services**

Our proposal includes developing and delivering a number of services to the public health and wider earth science community, including software tools, data products, and packaged models.

**Software Tools:** We will provide the earth science community with a modeling platform capable of assisting the user in accessing federated datasets, identifying relevant covariates using spatial data mining techniques, pinpointing regions of interest using content-based queries methodologies, developing and validating models using statistical techniques. The modeling platform will consist of a set of components (JavaBeans), that can be individually downloaded and distributed across the federation.

**Data Products:** By relying on the InfoPyramid data representation, we will provide facilities to make data products accessible to the Federation. In particular, we will allow our user groups to publish data used to conduct their studies, as well as the outputs of the models.

**Packaged models:** We will be providing packaged epidemiological models developed with our tool to the public health community at large. Such models

will be downloadable by the user and applied to data sets other than those used for model development.

In summary, we intend to build a prototype modeling system which employs digital images from EOSDIS as well as a variety of other data sources, and which supports interoperability of both data and software components. We propose making the modeling system, search engine, raw data, and modeling output available via the Internet to other ESIP partners and research communities. The testbed will examine a data collection that is of significant size, and will be used to estimate the quality of the techniques when applied to databases much larger than the testbed database.

## **1.2 Project Outline**

The proposed effort is a partnership, under the guidance of Dr. Chung-Sheng Li of IBM T. J. Watson Research Center, between the IBM T.J. Watson Research Center and the Johns Hopkins University (JHU) Program on Health Effects of Global Environmental Change. The focus of this effort is to develop a testbed which demonstrates that advances in epidemiologic analyses can be combined with content-based image and data retrieval techniques. The goal is to address extensive, large-scale problems of public health concern by merging remotely sensed (and other geographically related) data with epidemiologic analysis methods. Through a series of six case studies, we propose to make fundamental contributions to both the methodology of remotely sensed data analysis in an integrated application setting and in the applications themselves. The aforementioned six studies involve: Hantavirus Pulmonary Syndrome, Lyme disease, fire ant infestations and control, urban air quality and malaria (2 studies). These case studies were selected to characterize different aspects of problem solving and decision making for public health researchers and policy makers that could be addressed by the innovative use of remotely sensed and other large spatial data bases. The major approaches for each scenario are sketched out under each section. Details of the proposed analytical approach are given for one example (Section 1.4).

Many enabling technologies are already in place to support this project. The first contributor is the ongoing project at IBM T. J. Watson Research Center on *Retrieval of Digital Images by Means of Content* funded in part by NASA/CAN NCC5-101, to be completed by the end of this year. In this project, a progressive framework is developed to allow scaleable image and data retrieval (Li et al., 1997). Furthermore, an extensible query framework is established to allow the user to compose complex entities based on semantics, features and pixels (see section 1.9 for a more detailed description). The second contributor is the epidemiologic risk assessment conducted at the Johns Hopkins University School of Public Health Program on Health Effects of Global Environmental Change (JHU PHEGEC). This program involves various active researchers within the University who collaborate with a number of federal agencies on public health problems related to environmental conditions, both anthropogenically induced, as well as natural varying. Agencies currently involved include: the Centers for Disease Control & Prevention, the National Institutes of Health, the Indian Health Service, the U.S. Department of Agriculture, and the U. S. Environmental Protection Agency. These

agencies recognize the need to extend currently funded projects and will support the use of the epidemiologic data gathered as part of these studies in the current CAN.

From a methodological perspective, we see our project as addressing aspects of public health concern relevant to EOSDIS and WP-Federation that have two particularly challenging attributes: (1) the remotely sensed data sets with which we will work can be very large, hence efficiency of the algorithms we will use to analyze the data is a critical concern; and, (2) our speculation is that the relevant features we will need to extract from the data are often spatially very complicated. The resulting challenges are magnified in that the data and processes that we intend to study are dynamic. When the system is developed, the flexibility of the design will make archival remotely sensed data invaluable for retrospective analyses of public health problems that have not been amenable to previous analyses.

Our basic hypothesis in pursuing this study is that from a public health perspective, there exists a largely untapped wealth of highly relevant remotely sensed data. However, this requires improved mathematical and visualization tools to make the extraction and analysis of this data feasible, enabling its ultimate use in policy analysis and support.

We propose a dual-thrust approach in which we will first conduct our analyses focusing on current and past data. Then, for selected applications we will use the developed models to predict current patterns of disease and validate these models using recently produced images from EOS.

Commercialization of the research in the future is a significant possibility. IBM is currently engaging in various business activities related to the research to be conducted in this proposal, including digital libraries, database extenders, and multimedia miners. While there is no commitment from IBM to establish a business based on the research results, the research will be evaluated internally to determine if such a business is a worthwhile venture for IBM. IBM is prepared to contribute 50% of the project costs to meet the cost sharing requirements when a project has potential for commercial exploitation.

### 1.3 Proposed Scenarios

*(1) Predicting Hantavirus Pulmonary Syndrome (HPS) outbreaks – This scenario will demonstrate the application of this system to predicting environmental conditions associated with infectious disease outbreaks with sufficient lead time that public health interventions can be instituted. The data gathered resulted from an epidemiologic investigation of the HPS outbreak in the U.S. Southwest in 1993-94 and is supported by CDC and IHS.*

Hantaviruses are maintained in a select group of rodents around the world. When people are infected, the results vary from mild flu-like illness to death, depending on several factors including the strain of virus. In the spring and summer of 1993, there was outbreak of acute respiratory distress with a high fatality rate (> 70%) among previously healthy individuals. Mortality was traced to infection with a previously unrecognized hantavirus (Nichol et al., 1993) that was transmitted by a common native field rodent, the deer mouse (*Peromyscus maniculatus*) (Childs et al., 1994). This discovery was

significant because the rodent is widely distributed throughout North America, ranging from the Arctic circle to the Mexican Plateau and from the Pacific Ocean to the Eastern U.S. Surveys show that the virus is widespread throughout the range of the rodent so that most individuals residing in rural areas of North America are potentially at risk for HPS. Consequently, the question arose as to why outbreaks had not been previously recognized. Recent work suggests that climatic conditions associated with an El Nino event led to conditions favoring increased rodent populations through the spring of 1993 so that rodent populations in the area of the outbreak reached record levels (Parmenter et al., 1995). This was thought to be due to the above-average precipitation patterns in the region, leading to increased rodent survival and reproduction during the summer and autumn of 1992, a time when rodent populations usually decline in response to arid conditions. An epidemiologic analysis of the residences of households associated with HPS showed that these sites were associated with higher abundance of infected deer mice than houses where there were no cases of HPS (Childs et al., 1995).

The integration of remotely sensed data and other large data bases is critical for examining human risk for HPS for several reasons. First, the size of the geographic area involved (> 180,000 sq km in the original outbreak region) and the scattered rural population makes it unfeasible to use standard methods to identify the population at risk. Second, because of the habitat requirements of the deer mouse, not all areas are equally at risk. Third, as indicated above, although the deer mice always occur within the region, the timing and sequence of climatic conditions are critical to producing conditions suitable for an epidemic.

We will use the content-based retrieval techniques outlined in the next section (Section 1.4) to examine these hypotheses by building a temporally dynamic analysis of factors associated with increased HPS risk, with the goal of predicting HPS outbreaks in the U.S. Southwest. Such an approach is needed because of the apparent dynamic process of environmental conditions that produce situations favorable for HPS outbreaks. As a first step, the current epidemiologic data (case and random addresses) will be used to train the system to identify conditions associated with temporal trends in reflectance patterns from TM data during the winter - summer of 1991-92 to develop a classification algorithm for sites at risk for HPS. The classification algorithm will be validated by applying it to previously unclassified residences in a more extensive region during the same time period. Next, we will evaluate the algorithm by attempting to predict past outbreaks of HPS that were not recognized by public health researchers. We will use the developed decision rule to search remotely sensed images prior to 1992 for times and places where conditions for HPS were predicted to be favorable. Using our current collaborations with IHS and CDC, we will access state morbidity and mortality records for those times and places for cases of death compatible with HPS. Contacts with state health departments will be performed for those cases to determine if post-mortem materials remains. These remains (usually sera or tissues) will be tested in our laboratory using standard methods. Previous surveys of post-mortem archival materials dating more than 15 years prior to the recognition of HPS have been conducted. Although these studies lacked environmental analyses to focus their specimen selection they found that at least 15% of patients with symptoms compatible with HPS had died from hantavirus

infection (Zaki et al., 1996). With an environmental model of risk factors we anticipate a significant improvement in the predictive power of case detection.

If this model is developed and applied, it may be possible to identify where and when outbreaks of HPS are likely to occur with sufficient lead time (up to 1 year, based on our preliminary analyses) that public health interventions could be implemented (Glass et al., 1997). As such, it would serve as a model of early warning system studies for disease outbreaks that incorporate remotely sensed data over large geographic areas. Recently, NOAA has issued a warning to public health services that a significant ENSO event is being forecast which may have potential implications for HPS.

**(2) Predicting Lyme disease (LD) dynamics – This scenario will demonstrate the application of this system by integrating data from various sources to characterize annual levels of variation in a well-established vector-borne disease. The epidemiologic data to be used in the study was gathered under studies supported by CDC and NIH.**

Lyme disease (LD) is the most common vector-borne disease in the U.S. with more than 14,000 cases reported in 1996 to the Centers for Disease Control and Prevention (CDC). LD is caused by a spirochete, *Borrelia burgdorferi*, that is transmitted from animals to people by the bite of a tick. The most commonly associated tick in the eastern half of the U.S. is *Ixodes scapularis*, the black legged tick. Several species of mammals play an important role in LD maintenance either by harboring the spirochete (e.g. white-footed mouse) or by serving as food sources and mating sites for adult ticks (white-tailed deer). Along the East coast, most human cases occur from northern Virginia through New England.

LD is difficult to diagnose based on currently available clinical characteristics and diagnostic tests. In addition, the numbers of LD cases vary dramatically in both time and space throughout the region. For example Baltimore County, Maryland reported 92 cases of LD in 1996, while Carroll County, Maryland which is immediately west, reported only 27 cases. Although various factors influence the observed numbers of LD cases, we have shown that environmental covariates related to conditions affecting tick and animal survival and abundance are good predictors of both adult tick abundance (Glass et al., 1994) and proportional human disease risk (Glass et al., 1995). Identification of the population at risk is critical to anticipating public health intervention needs, as well as planning treatment and vaccine delivery programs, which will be developed following U.S. FDA approval of currently tested vaccines.

Thus, the detailed, overall geographic risk of LD for the eastern U.S. is critical to assessing the population at risk in this region. In addition, patterns of precipitation and temperature throughout the year affect the absolute numbers of ticks and animal reservoirs in suitable habitat by influencing survival and reproductive success.

Preliminary studies show currently available data related to soils, elevation, slope and aspect, watersheds and land use patterns can predict patterns associated with tick abundance and human disease risk (Glass et al., 1995). These studies rely on the analysis of patterns of spatial covariation of environmental conditions to predict outcomes. To predict absolute levels of human disease risk which are influenced by vector and animal

abundance, however, we will need to incorporate RS data that provide indicators of environmental conditions in the region of study.

The studied region will include an epidemiologic data base already developed for Maryland, Virginia, Delaware and eastern Pennsylvania. This includes information on the locations of human LD cases from 1992 to the present. These data incorporate the time of diagnosis (as an estimate of time of infection), and presumed place of exposure. These time-location data will be used as the outcome variables to be measured. Because of the marked annual periodicity to LD, we will model the between-years variation in the rates of infection, locations and the timing. Environmental data based on soils, elevation, slope, aspect, hydrologic features will be incorporated as baseline predictor variables. Information on temporal patterns of temperature and precipitation for the region will be acquired from NOAA, and Landsat TM or MSS data will be used to characterize variation in soil moisture and vegetative growth in the time periods preceding LD cases. The goal will be to generate a climate driven model that predicts deviations of LD cases from baseline for the region. Results will be validated by using the query system to predict the temporal-spatial patterns of LD in the granting period (1998-2000) and comparing them with results reported to CDC, the federal agency responsible for national reporting of the disease.

(3) *Application of remote sensed and spatial data for agricultural and public health impact of fire ants in the United States – This scenario will examine the ability of the system to predict distribution and spread of introduced pest species and to model the effects of proposed control measures on the pest species. The entomological data are provided by various state agencies and the U.S. Department of Agriculture.*

The red imported fire ant, *Solenopsis invicta*, was introduced into the United States from South America about 50 years ago (Vinson and Sorensen, 1986). Today this exotic pest infests close to 250 million acres in the southeastern United States. Imported fire ants have no natural enemies and adapt to changing climatic and environmental conditions. It is estimated they could infest almost a quarter of the land mass of the United States. High densities of *S. invicta* cause numerous environmental and economic problems (Lofgren, 1986). In urban areas ants infest yards, playgrounds and open fields, and as many as 200 mounds per acre have been documented. During flooding, water borne ants sting on contact.

In the southeastern United States, fire ants are the number one cause of known stings and bites. In infested urban areas, 30 to 60% of human population are stung every year, with children being most affected. Between 17 and 56% of persons stung experience large local reactions at the site of the sting (deShazo et al., 1990). In extreme cases, swelling may require treatment with steroids, antihistamine, decompressive surgery or amputation. Up to 2% of stings result in life threatening anaphylaxis reaction (Stafford, 1989). Morbidity due to secondary bacterial infections is as high as 54% (Triplett, 1976), because the sting site is intensely pruritic.

In addition to the toll to human health from envenomization, anaphylaxis, and secondary infections (Stafford et al., 1989), fire ants attack livestock and damage crops, infrastructure (roadway collapse), and telecommunication electric insulation. The

estimated economic impact of fire ants in Texas is about \$111 million annually (Lofgren, 1986; Vinson, 1997). Additionally, there are the costs associated with mandated quarantine (Lofgren, 1989), and the growing realization that fire ants have a major impact on biodiversity and resource conservation that can only be addressed in area-wide management strategies (Allen et al., 1994). Although a number of baits have been developed for homeowners, they are not cost effective, nor environmentally sound for use in large areas. High reproductive and dispersal rates ensure reinfestation and expansion of this species.

The success of future control technologies will be largely dependent on accurate and timely targeting of interventions on a spatial and temporal basis. Many critical parameters for targeting include variables compatible with remote sensing, such as soil moisture and temperature, precipitation, topography and plant communities, and atmospheric moisture. These govern distributions of fire ants in both time and space, and weather factors influence daily and seasonal behaviors. Therefore, it is imperative to develop a sophisticated remote sensing-based system that can provide locally-relevant, near real-time, decision support technology. Information ultimately must be web-resident to facilitate access by local end-users. We anticipate the final system will integrate previous and current conditions relevant to seasonal dynamics of the populations on a spatial basis. Using national data bases of fire ant distribution provided by Imported Fire Ant & Household Insects Research Unit, Gainesville, Florida, we will use the IBM system to: (1) develop near-real time algorithms to identify risks from fire ant activity; (2) develop algorithms to facilitate spatio-temporal targeting of interventions to maximize biological impact; (3) develop algorithms to use remote sensing data to predict potential areas of spread in the U.S.; (4) use remote-sensing data to match weather and environmental characteristics of sites with candidate biological control agents from South America to potential release sites in the U.S.

*(4) Air Pollution and Health Related Effects – This scenario will be used to study the use of remotely sensed data to examine the impact of air quality on human health outcomes in urban areas. Epidemiologic data for these studies are derived from studies from the U.S. EPA and the Health Effects Institute.*

Ecological impacts and human morbidity and mortality from criteria air pollutants has been well-documented (Bascom et al., 1996; Samet et al., 1995a). Modeling pollutant loads (Ellis, 1994) using multiple regression techniques which account for both climatic and pollutant exposures (Samet, 1995b) have been developed by the Hopkins co-investigators. Under current funding by the U. S. EPA and the Health Effects Institute, air pollution databases are being constructed for over 100 U.S. cities. We propose to investigate new opportunities presented by this CAN to utilize remotely sensed data in the modeling and analysis of selected air pollutants - with a special emphasis on urban air quality. Pollutants include, for example, particulate matter, tropospheric ozone and SO<sub>2</sub>. A unique signature of the work will be our attempt to map emissions of ambient concentrations and link them to selected epidemiological end points for these U.S. sites - all enhanced through the efficient use of remotely sensed data.

The goal is to integrate remote sensed data, mesoscale modeling, mortality and hospital admission data from the National Center of Health Statistics, and newly

developed data analysis tools, and thereby gain a better understanding of existing urban air quality and cost-effective means for its improvement. The connections between remotely sensed data and air quality modeling and analysis differ considerably for the pollutants listed above. The most obvious connection can be made for tropospheric ozone, which, in part, is driven by temperature. The connections are less obvious for SO<sub>2</sub> and particulate matter.

The major components in the modeling and analysis we propose are: establish emissions inventories (Ellis et al., 1994); acquire and process meteorologic data; implement several mesoscale models, and acquire and process epidemiologic data. Remotely sensed data will play a central role in building emissions inventories (Luman and Ji, 1995), calibrating and validating mesoscale modeling efforts and augmenting epidemiologic information. Landsat-5 TM data will be used to help characterize historical patterns of anthropogenic and biogenic VOC emissions for subsequent use in a regional ozone production and transport model. These data then will be used to provide more accurate estimates of timing and durations of exposures to air pollutants, and to examine their associations with morbidity and mortality in urban populations.

Initial efforts will focus on urban nonattainment areas for tropospheric ozone in the Philadelphia - Baltimore - Washington corridor. After initial model development, validation efforts will focus on modeling the same phenomenon in Los Angeles, Denver and Chicago. An interesting and productive possibility would be to design studies so as to make use of existing work reported in the recent (April 1997 draft) EPA report: *The Benefits and Costs of the Clean Air Act, 1970-1990*.

**(5) Forecasting malaria in India using climate cycles, remote sensing and GIS – This scenario will demonstrate the application of this system in forecasting disease risk in which both natural climatic variability and land use changes interact to significantly impact risk of emerging diseases.**

Malaria is the most prevalent vector-borne disease globally and causes 1-2 million deaths annually. A marked resurgence of malaria has occurred in the last 20 years (IOM, 1991). Malaria has reemerged in India since the 1980s and today represents one of India's most significant public health problems, with an estimated 141 million people living in epidemic prone areas (Sharma, 1995). Research to investigate potentially improved forecasting is, therefore, a high priority. The goal of this scenario is to identify climate and land use-related risk factors for malaria epidemics in India, by applying RS image analysis to a long time series of high quality epidemiological and environmental databases.

The epidemiology of malaria depends on the interplay of many variables. Changes in climatic conditions affect the longevity as well as the infectivity of malaria-transmitting mosquitoes (Gilles, 1993). Recently (Bouma et.al., 1994, 1995, 1996) have demonstrated the importance of El Niño Southern Oscillation (ENSO) in predicting the monsoon rainfall in India and its relation to the malaria epidemics. Although El Niño predicts overall rainfall, it is not efficient at predicting rainfall at particular sites where risk of malaria needs to be calculated.

Satellite imagery, coupled with El Niño climate forecasting, needs to be integrated to obtain information pertinent to the multiple etiologic factors leading to malaria

epidemics. These factors include urbanization, water resource development, detailed landuse and irrigation patterns, vegetation changes, soil moisture and terrain characteristics. India's irrigation system, for example, is extensive and has changed the ecology and human population density over vast areas. This has created an ideal situation to study the combined effect of agricultural development, population migration and climatic variation on malaria incidence and prevalence in a location with high quality human epidemiological data.

The study area will comprise the Gujarat and Rajasthan provinces. While climate/El Niño correlation is expected to explain much of malaria variation in semiarid Rajasthan, a more sophisticated analysis of irrigation and associated factors will require sophisticated RS software for analysis. The disease variables used in the study are malaria incidence and blood-slide positivity rate. Monthly morbidity data (microscopically confirmed) is available from 1960-1995 at the district level, and data on use of pesticides is readily available. Ancillary environmental data, including annual maps of the irrigation networks are available since 1950, and Dr. Bouma has rasterized climate data for the region from 1961-1990 provided by the Climate Research Unit, University of East Anglia. These data will serve as a long term record to augment potentially limited time series data from district databases.

Biological and climatic, agricultural, water management, vector control activities and demographic variables will be used to train the IBM software to associate changing patterns of malaria incidence with measured predictor variables. Both regional and district level analyses will be based on monthly malaria case data for the period 1960-1995. Data from NOAA AVHRR and METEOSAT, which can provide information on spatial and temporal variations in environmental parameters will be used to characterize local impacts of ENSO activities.

**(6) Malaria associated with deforestation in Peru – This scenario will be used to study the impact of land use changes in areas surrounding urban areas in developing countries on patterns of infectious diseases. The epidemiologic studies have been supported by NIH and the U.S. Department of the Army.**

One major element of malaria surveillance and proactive health intervention involves monitoring the populations of mosquito vectors, in relation to human settlements. Use of remote sensing and spatial analysis has been shown to be useful in identifying environmental factors determining the temporal and spatial distribution of mosquito breeding sites (Beck, 1994; Marques, 1987; Washino & Wood, 1994).

Extensive deforestation around the jungle city of Iquitos, Peru, provides an excellent ongoing ecological experiment in the relationship between landuse change in urban areas of developing countries and disease. In the last three years there has been an epidemic spread of malaria in the areas contiguous to Iquitos. This explosive upsurge in malaria coincides with heavy deforestation, and the subsequent influx of *Anopheles darlingi* mosquito, rarely found in the area previously. Now, this malaria vector constitutes 97% of the local mosquito population.

The malaria epidemic in Iquitos occurred in one of the few jungle areas of Peru where there is a strong data base of human and insect vector surveillance. The Ministry of Health has followed the course of both *P. falciparum* and *P. vivax* for the last ten

years. The increase in cases and their location is well documented in several different Ministry of Health data bases. At the same time the U.S. Naval Medical Research Unit (NAMRU) entomological station has been performing studies on the insect vectors that had been present in the area before the current epidemic. Records for the last ten years on temperature and rainfall are available from the Peruvian weather stations located in and around Iquitos.

Socioeconomic and environmental changes associated with progressive stages of deforestation and human settlement are suitable for study by remote sensing (Green & Sussman, 1990) We will investigate the relationship between deforestation, peri-urban development and malaria after 1987. We will employ satellite images from 1987 onward to document qualitative and quantitative landcover change, correlate it with changes in mosquito population composition and to the emergence of a significant increase in human malaria cases. We hypothesize that by using interpretive software developed by IBM applied to NASA satellite images, we can better characterize the sequential changes in malaria risk associated with deforestation and build predictive models of malaria risk based on the patterns of ecological change. These results then can be applied to other rapidly developing urban areas in tropical areas with the goal of characterizing changing patterns of disease risk.

## **1.4 Content-based Retrieval for Model Generation and Validation**

### **1.4.1 Enabling Technology**

An image navigation and content-based search engine has been developed by IBM under NASA CAN NCC5-101 (Li et al., 1997). The resulting internet-based client-server system supports image browsing and navigation, content-based query, and visualization. The primary technology focuses of this project have been content-based query specification and search techniques, with particular regard to efficient storage and retrieval of satellite image data.

Our technology is based on the notion of progressive layerings of information, or information pyramids. There are two main sets of such pyramids, abstraction pyramids, for representing searchable content, and resolution pyramids, for supporting multi-resolution image retrieval and processing.

The resulting search framework, is based on a technique, Multiple Abstraction Level Content Based Retrieval (MALCBR, Castelli et al. ,1997) in which each search target is defined in an object-oriented fashion. An atomic object can be defined at the semantic, feature (e.g. texture), or pixel level of the abstraction pyramid. A composite object consists of multiple atomic objects with spatial or boolean relationships. For instance a composite object might consist of two houses and a road, where the houses are separated by no more than 100 meter and are within 50 meters of a road. Figure 3 shows the results of a search where the houses are framed by a solid line and the road by a dashed line.

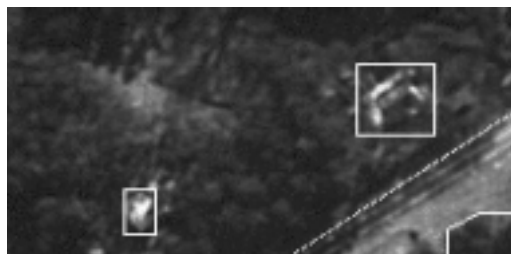


Figure 3: Composite object consisting of two houses within 100 meters of a road.

The abstraction pyramid enhances the efficiency of content-based search by allowing the search to proceed on low-volume data abstracts when available, while preserving the flexibility of searching at lower levels when predefined semantics or features do not adequately capture the user's intent. We achieve additional efficiency in processing compound object definitions by using graph-search and dynamic programming algorithms.

Image retrieval at multiple spatial resolutions is supported by means of a multi-resolution pyramid. A novel wavelet-based compression scheme is used to efficiently store and access images or portions of images at a variety of scales. Using this multi-resolution scheme, we have achieved notable speedups (anywhere between 10 to 30 times on average) on a number of image processing operators used for content-based retrieval, including classification and template matching.

An additional component of the system is the image navigation client. This is a Java applet which supports metadata browsing, map views, image selection, multi-level zoom, multi-layer vector overlays, and gazeteer functions. Content-based queries are composed using a novel drag-and-drop interface (DanDE, Bergman et al., 1997) that allows the user to create English-language-like statements that describe simple and compound objects as well as query constraints. DanDE (short for Drag-and-Drop English) is a programmable interface builder that allows for rapid prototyping of new query interfaces. Figure 3 is an example of creating a content-based query using DanDE.



Figure 3. Sample content-based query using DanDE

### 1.4.2 Proposed Solution

We propose a new methodology to provide a common platform for combining content-based retrieval, spatial and time series data mining, and statistical modeling. In this proposal, major extensions to MALCBR are proposed to support the scenarios in the previous section. Specifically,

1. We will extend the object definition capabilities in order to encapsulate salient spatial features (such as texture and contour), temporal features (moving average, slope, length), spectral features (such as spectral histogram), and statistical features (such as moments, skew). Each object can be viewed as a filter to screen the candidates in the database (pre-extracted or extracted on the fly). The processing of each object will produce a ranked list based on a given similarity measure. We propose to adopt an extensible definition framework for earth science data, by developing a new specification language for manipulating objects, random variables and time series.
2. We will extend the object composition capabilities to accommodate spatial, temporal, spectral, and regression relationships. This object composition capability is essential for constructing composite objects that contains spatiotemporal constraints among components of the object. In particular, we will focus on novel techniques for establishing constraints, provide hints to the spatial data mining for association rule discovery, and assist statistical modeling components to build regression models.
3. We will establish the capability to adaptively adjust the similarity measures used in content-based search through iterative refinement. We propose to use a modified version of nonlinear multidimensional scaling to dynamically “warp” the feature space based on user feedback and the historical database.
4. We propose a new technique, to efficiently parse and process a composite object to eliminate the possibility of false dismissal. This technique processes bounded rules (such as the definition of water) first, then fuzzy rules (such as the similarity in the texture space), then unrestricted rules (such as positional constraints on pairs of objects).
5. We propose a new technique to reduce the dimensionality of the features extracted from the image without reducing the precision and recall of similarity retrieval. This technique, called RCSVD, applies clustering and singular value decomposition recursively to minimize the total amount of storage space required to store the feature vectors extracted from the images, time series, or other data sets and to use very efficient low-dimensionality indexing methods on the resulting projected data. The benefits are twofold: by reducing the volume of data, RCSVD utilizes the computer memory hierarchy more efficiently than existing schemes, and by reducing the dimensionality of the search space it allows to use off-the-shelf algorithms such as R-Trees, which are highly optimized for low-dimensional search spaces but perform extremely poorly on high-dimensional datasets. Initial results show that this approach is a promising solution to the scalability issues arising from searching large earth science databases.
6. We propose to extend the "plugboard" approach originally developed in the NASA CAN, whereby tools can be easily invoked and new tools created to

provide query, modeling, and data mining capabilities. The new tools to be integrated include spatial data mining tools as well as statistical tools which allow the user to build regression models, to test hypotheses, and to discover spatial and temporal associating rules. This approach takes advantage of the emerging technology “java beans” and provides an intuitive and interactive mechanism for the user to construct new data processing flow, new objects (both atomic and composite), and new models. An object-oriented approach is adopted to allow data to be passed or shared among different components.

7. We propose to extend our current progressive framework to spatial data mining algorithms to further improve its scalability. This will be accomplished by combining data representation and data mining, in the spirit of NASA CAN NCC5-101. In particular, progressively represented spatial data allows analysis techniques to be applied hierarchically and thus further performance improvements can be expected over the state of the art.

### **1.4.3 An Example**

We illustrate the proposed approach through means of an example. Consider developing a model to predict outbreak of Hantavirus Pulmonary Syndrom using environmental factors as predictors, described in Section 1.3.

We can model virus outbreaks by locating rural houses surrounded by appropriate vegetative cover, and performing an analysis in these areas. The analysis will be based on vegetation derived from Landsat TM imagery, soil moisture from AVHRR imagery, weather patterns from NWS ground stations, and epidemiological records.

Model development relies on data from a limited spatial region for which good epidemiological data exists. Using the federated search capabilities accessible from our system, the user will be able to identify potentially relevant datasets, and to specify the region of interest and time frame for the model development activity. This identification phase is likely to take the form of a dialog in which the user specifies terms of interest (e.g., “precipitation”, “soil moisture”) and other metadata to a remote request manager which will return catalog and explanatory information, to be used to “zero in” on the desired data products.

The regions of interest for the study are defined as areas immediately surrounding isolated houses. Let us assume that the user has no knowledge of algorithms for identifying isolated houses from medium resolution datasets, such as Landsat TM imagery. The user, nevertheless, knows the exact location of the houses in the limited region for which epidemiological data has been collected. The user guesses that isolated houses could be located by means of texture features, spectral signatures, and/or digital elevation maps available for the region, and consequently selects a large number of descriptors. The system cannot train a traditional classifier, because the number of inputs triggers the so-called “curse of dimensionality” (i.e., the classifier would require a number of training examples that grows exponentially with the number of input parameters). It is necessary, therefore, to reduce the number of inputs to the classifier. This can be accomplished with knowledge discovery techniques supported by the system, such as attribute focusing. Isolated houses detected by the classifier are the basic search composite objects (that is, they are the composition of the atomic objects *house* and

*barren* terrain through the composition operator *surrounded by*) used in the study, and they form a mask for other spatial and temporal variables to be used in the model.

The user must now select candidate predictors for the study. In particular, the user selects average temperature, precipitation and soil moisture indices as predictors to identify the macro-climatological conditions that seem to favor the outbreaks of the virus. If some of the predictors are not directly available in federated datasets, for instance, the soil moisture index, the system allows the user to specify how to derive such quantities from other data products, for instance, how a soil moisture index can be derived from AVHRR.

A first analysis, performed on these large-scale predictors, shows a strong dependency of the outbreak of the virus on unusually wet conditions, such as those associated with the presence of El Niño. The user then proceeds to specify a set of dependent values for the identification of local conditions that promote the actual outbreak of the disease. The selected variables are the reflectance in the six reflective bands of TM averaged over a small window centered on the location of the house (to reduce sensor noise and to account for possible misregistration of the satellite imagery), the time series (including lagged values) of the local value of vegetation index, weather data, and soil moisture near the locations of the houses, at user-specified time intervals and for a user-specified time period.

The user now selects a type of model from several predefined types; for instance, a linear regression model or a nonparametric method. The system fits the model to the data, with epidemiological data being used to define the response. The system provides the user with the relevance of the different predictors to the model, and the user discards those predictors that appear irrelevant to the model. The model is then retrained and alternate families of models are considered, until a satisfactory solution is generated through iterative refinements. The scalability of the resulting model to a large dataset is ensured by the MALCBB and RCSVD techniques discussed in the previous sections.

Note that a variety of operator types are used in formulating this query. Much of the user interaction is example-based. Features such as those used in defining the “house” search object will be selected by the system based on user-provided locations of known houses. The user will receive feedback from the system to assist in the specification of the query, and visualization techniques will be applied extensively. An interface which allows hierarchical compositing of features and relationships will be based on our current DanDE (Drag-and-drop-english) interface.

The user now validates the model by specifying a different year for which epidemiological records are available. The system keeps track of the datasets used for the initial analysis as part of the model definition, and retrieves relevant portions of this data for the validation, either locally, or from the federation. The user may use a variety of visualization tools to view results of the model run on the validation set, and compare results with the initial development run.

Once validated, the user chooses to apply the model to another geographic region. InfoPyramid, discussed in Sections 1.5 and 1.6, will facilitate the selection of alternate datasets. For example, Landsat TM data may not be available for the new region. The system will receive suggestions from InfoPyramid, indicating that vegetation indices from AVHRR are available for the new region, and might be used as stand-ins for the Landsat

TM-derived vegetation indices. The user will run through a set of procedures for recalibrating the model with the replacement dataset and then receive results for the new study area.

As a final step, the user now decides to create a stand-alone application from this model. A framework will be developed to permit connection of modeling components to a set of JavaBeans UI components in order to package the model as a distributable application.

## 1.5 System Architecture

We design a system architecture which coordinates the search and retrieval of spatial data and earth observation images from distributed, federated data archives. We are proposing two novel technologies, the InfoPyramid and the SFGraph, and combine them with the Z39.50 Information Retrieval protocol and Catalog Interoperability Protocol (CIP) retrieval profile in order to provide a complete and powerful system for the search, retrieval, manipulation and delivery of spatial data and earth observation images.

The InfoPyramid provides a framework for integrating data and methods for retrieving and generating data and data abstractions, translations and substitutions at the time of the search. More specifically, InfoPyramid provides a schema to specify and integrate data from different instruments (e.g. TM, AVHRR, MODIS), different resolutions (30 m or 1100m), different spectra (from microwave to deep infrared), different time, and different abstractions (e.g. texture features, spectral histogram, and semantics of the image). The wide varieties of data sources for the same location at a given time correspond to different *modalities* of that location. The SFGraph provides a framework for the progressive delivery and storage of large volumes of data. The Z39.50 protocol (Z39.50, 1995) provides a basic framework for the search and retrieval of data records. CIP provides a framework for achieving catalog interoperability through cooperative independent Retrieval Managers (CIP, 1997).

We propose that the CIP system be utilized to coordinate the search and retrieval of both the data and the methods from the distributed, federated sources. In response to client searches, the Retrieval Managers construct transient InfoPyramid structures by retrieving or generating the requested data depending on the availability of source data and data generation methods. The Retrieval Managers can directly follow through on executing the data generation methods, or can package the methods and necessary data for execution at the client.

The architecture of the proposed working prototype, as shown in Figure 1, consists of three component classes: Data and method servers, Retrieval Managers and World-Wide Web- and Internet-based clients.

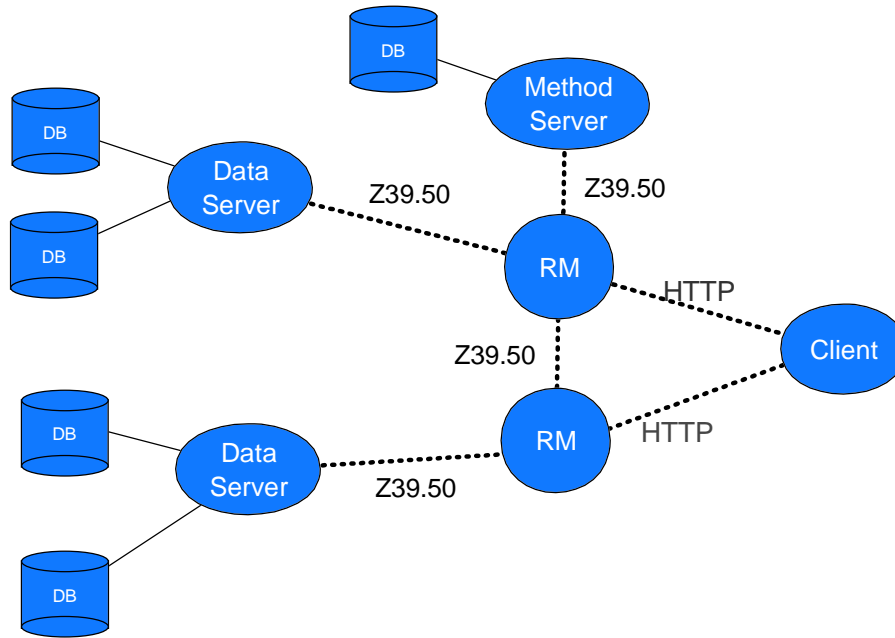


Figure 1. Proposed working prototype architecture which combines InfoPyramidal data representation, Z39.50 search and retrieval protocol and CIP Retrieval Managers (RM).

*(1) Internet-based Clients*

The Internet-based clients are capable of generating queries, retrieving data and displaying query results. The lightweight clients are developed in Java and are platform independent. The user downloads the client by selecting a particular RM in the netscape or explorer browser. Consequently, the client has affinity to a specific RM once a client session is established. Data providers participating in the federation expose their data archives through Retrieval Managers. Clients communicate with the data archives through these Retrieval Managers, which carry out the client metasearches and merge the retrieval results for presentation to the clients.

*(2) Retrieval Managers*

Similar to the Retrieval Manager concept in the CIP specification, the data providers make the data archives available to the federation by implementing compliant Retrieval Managers. Since the Retrieval Managers are capable of communicating with each other, the mechanisms of federated search are transparent to the client. Furthermore, since the Retrieval Managers are automatically detected by automated agents that search the Web, the implementation of a Retrieval Manager is sufficient for exposing the corresponding data archives.

We extend the capabilities of the Retrieval Managers in order to allow query-time data abstraction, modality translation and substitution. These enhanced functionalities are critical in the processes of generating, validating and applying scientific models involving spatially- and temporally-sensitive data and earth observation imagery. In general, the manipulation of data in the search and retrieval process is accomplished through methods which are retrieved from on-line method archives. For example, these archives provide

methods for reducing the resolution of images, extracting features, segmenting images, classifying data, combining data modalities, reducing noise, and so forth.

The Retrieval Managers communicate the queries and retrieve the results via the Z39.50 protocol. By combining the method services with the data sources, the Retrieval Managers enhance the functionality of the Z39.50 protocol by operating on and manipulating the data before delivery to the client. This enhanced functionality is not supplied by Z39.50 or the base CIP system.

### *(3) Data and Method Sources*

The data archives are exposed via the Retrieval Managers through access points. The access points define the attributes of the data records upon which searches may be conducted. Typically, the access points for the earth observation imagery are spatial and temporal coordinates, spectral bands, and so forth. We introduce additional access points, such as abstraction level -- resolution, features, and so forth, and modality, which allow for the generation and manipulation of the data via the InfoPyramids. The method archives are treated similarly to the data archives. The method archives are published via Retrieval Managers, and the method records are searched via the abstraction level and modality access points.

## **1.6 Innovative approaches for data production.**

We describe the innovations provided by the InfoPyramid and SFGGraph in the representation, storage, retrieval, delivery and distribution of data.

### *(1) Data Representation -- InfoPyramid*

The InfoPyramid provides a new and powerful system for data abstraction, data substitution, delivery and interoperability within the federated environment. In providing methods for data abstraction, the InfoPyramid allows delivery of the specific aspects of the content which are relevant to the particular model or context.

For example, the data abstractions may consist of lower resolution versions of images, features extracted from image data, or summaries of the content such as classification results, and so forth. In providing methods for data substitution, the InfoPyramid encapsulates methods for automatically filling-in missing data. Since the InfoPyramid encapsulates both data and the methods for data generation, the InfoPyramid provides a wrapper for delivering data. The InfoPyramid data and methods are compatible with the Z39.50 information retrieval standard which supports interoperability within a federated system of distributed data stores and clients.

The InfoPyramid represents data at multiple levels of abstraction and in multiple modalities. The InfoPyramid fully encapsulates the data and the methods for generating new levels of abstraction and modalities into a single representation. In general, these methods are described by rule sets, which are fired in response to a request for data at a particular resolution and/or modality.

### *(2) Storage and Retrieval -- SFGGraph*

Enhancements in the efficient storage of the large volumes of data will be achieved by using a new method for the dynamic representation of

segmented/multi-resolution data based upon the Space and Frequency Graph (SFGraph). The SFGraph generates a compact representation of data, which can be adapted to attain various levels of compression, progressive access, hierarchical storage, and compatibility with user data access patterns.

For one, the SFGraph representation is perfectly suited for the storage of large satellite images. It decomposes the image into a hierarchy of spatial resolutions and spatial segments, which are stored as independent tiles. These tiles provide one set of views of the image, and can be assembled into a large multiplicity of additional views. By recording the history of user access to these tiles, the SFGraph representation can adapt the representation of the image, by migrating the tiles to various levels of storage in the storage hierarchy (i.e., tape, disk, memory). Since the full set of SFGraph tiles provides an over complete representation, the best subset of tiles can be dynamically selected to meet objectives of compression, view generation time, and retrieval-response time (Smith & Chang, 1997). In general, the SFGraph provides an efficient means for representing time series, spatial and volumetric data.

The SFGraph dynamic data representation provides enormous potential for cost savings in the storage, delivery and distribution of large volumes of science data. By adapting the representation to patterns of user access, greater efficiency is achieved in retrieving the typical views of the data. By decomposing the data into multiple resolutions and segments, the transfer of data over a computer network can be reduced by delivering only the required set of tiles that are necessary for constructing the user's requested view of the data. Furthermore, by dynamically and adaptively migrating the data representation to a multi-level storage hierarchy, storage costs can be reduced.

### *(3) Progressive Delivery and Distribution*

The efficient delivery of large volumes of data is achieved by progressive delivery using the SFGraph. In this way, the SFGraph is considered to be an instance of the InfoPyramid in which the data is stored at multiple resolutions and in a non-redundant fashion. For example, when a client requests image data at a particular resolution, the system constructs the InfoPyramid by packaging the appropriate SFGraph tiles and methods for combining the tiles in order to construct the requested image view.

The execution of methods for generating the requested data or image view may be carried out at the Retrieval Managers or at the client. When carried out at the Retrieval Managers, the generated data is retrieved directly to the client. When carried out at the client, the Retrieval Manager packages the required data and methods, and supplies these to the client. For example, consider the case of the delivery of SFGraph image data. The Retrieval Manager sends the client the SFGraph tile data and methods for filtering, sampling and composing the tiles that are required to generate the requested image view. The data and methods are packaged as a Java Bean which can be handled by any Java-based client.

## **1.7 Support for the Earth System Science Users**

The user group targeted for this proposed working prototype is the public health community, which needs to validate existing models as well as generate new models using both remotely sensed images and other environmental data sources. In order to

support this community, we provide Internet access to our search engine. The client is a java-based query builder, and thus runs on any platform with either Netscape navigator/communicator or Microsoft Internet Explorer. We will provide session management on the server side, to guarantee minimum level of quality of service for each user who is accessing the federated search engine.

The unique service provided to the user is a self-guided query and model building tool for interactively constructing and validating a hypothetical model. The tool is capable of iteratively refining the model and query through user feedback and access to a historical database. The unique product provided to the user is (1) output risk assessment maps, (2) the model extracted from the user hypothesis, feedback, and historical data, (3) the semantic and feature objects extracted from the raw data, (4) images which are progressively represented to facilitate fast mining, retrieval, and visualization.

The client guides the user in building a model and/or a query from those objects and features predefined by the system. This may include those well defined objects such as forest, grassland, water body. The user can also add additional definitions of objects and features to the system. These definitions can be made persistent for future reuse.

Since the client is Internet-based, the product and services are both available to the greater earth system science community.

## **1.8 Interoperability**

We propose interoperability within the federation via CIP Retrieval Managers (CIP, 1997) and the Z39.50 search and retrieval protocol standard (Z39.50, 1995). We describe several scenarios which demonstrate new levels of interoperability in the federated system.

*Scenario 1.* Data substitution. Consider that a scientist develops a model which requires precipitation data. Initially, the model is validated within a confined spatial area and time-frame. Later, the scientist applies the model to a larger spatial area and/or time-frame, in which the precipitation data is not fully available. The InfoPyramid provides a new level of interoperability by taking advantage of available data in the federation.

The client issues a search for precipitation data to a Retrieval Manager. The requested data is not found by the Retrieval Manager. The Retrieval Manager then issues a search to the methods store to request methods for generating the precipitation data.

*Solution 1.* Interpolation/extrapolation. One returned method generates the precipitation data for the required coverage by interpolating and extrapolating proximal data spatially and temporally. The Retrieval Manager searches now for the proximal data, retrieves the data, executes the method and returns the requested precipitation data to the user.

*Solution 2.* Modality Translation. Another retrieved method provides a formula for combining several modalities of earth observation images to produce a new data set which is highly correlated to precipitation data. The Retrieval Manager retrieves the images, executes the method and returns the requested precipitation data to the user.

**Scenario 2. Budget Constrained Search.** Further consider that there are a number of alternatives for retrieving and generating the data requested by the client. With each data product and data manipulation method is associated a cost of purchase, transaction and/or execution. Each alternate path provides a particular tradeoff between cost and quality. Given that the user has a limited budget, the Retrieval Managers choose between alternatives in order to maximize data quality for the user's budget.

In both these cases, the failure of the initial search, the retrieval of appropriate methods, the retrieval of alternative data and the generation of the requested data is handled automatically by the Retrieval Managers without requiring involvement of the user. This system takes greatest advantage of the redundancy in the data within the federation and the use of published methods for manipulating the data.

## **2. PARTICIPATION IN THE WP FEDERATION**

### **2.1 Federation Objectives**

- We will provide a framework for easy location, access and utilization of datasets from multiple ESIPS, based on the InfoPyramid approach.
- We will provide the earth science community with new modeling tools in support of emerging interdisciplinary fields, based on data from MTPE missions and global change studies.
- We will play a leading role in promoting new semantic interoperability paradigms for data product in addition to interoperability for metadata.

### **2.2 Participation Methodologies**

- *Architecture of the proposed WP ESIP:* our architecture is modeled on the one adopted by CIP, and is compliant to the all applicable FGDC standards.
- *User Interface:* our user interface will be based on Java (TM) and JavaBeans to allow cross-platform interoperability.
- *Cross-site search queries:* will be based on the Z39.50 Information Retrieval protocol layered on top of http. In the proposal architecture the retrieval manager acts as a metasearcher and prepares appropriate queries for the relevant WP ESIPS.
- *Data and metadata syntactical interoperability:* we propose InfoPyramid as a general framework for data representation, translation and abstraction. This object-oriented approach encapsulates both data formats and access methods to enable transparent access to multiple data and metadata formats and content.
- *Metadata semantic interoperability:* InfoPyramid provides the translation layers (in the form of profiles) to enable semantic interoperability at the metadata level.
- *Data semantic interoperability:* InfoPyramid provides interoperability at the feature level and high level data semantics.
- *Data Dissemination:* The architecture will support data dissemination through the Internet. In particular, the system will make accessible by other WP ESIPS in the federation as well as by the Internet community at large: 1) the (non-confidential) raw data used to formulate the models; 2) the products obtained from the validated models; 3) the models themselves, packaged as applications runnable at other WP ESIP sites.

## 2.3 Expected Contributions

- *Software Tools*: We will provide the earth science community with a modeling platform for accessing producing models based on federation data.
- *Data Products*: By relying on the InfoPyramid data representation, we will provide facilities to make data products accessible to the Federation.
- *Packaged models*: We will be providing packaged epidemiological models developed with our tool.

## 3. METRICS

We propose several metrics to measure our progress in contributing to the federation and for developing new technologies to support earth sciences.

### 3.1 Federation Metrics

We measure the contributions to the federation in terms of the extent of integration with the other sites in the federation and the progress in the development of the federated content-based search methods using the InfoPyramid.

- ***Integration with the Federation.*** To measure our progress in integrating with the federation functionally, we propose two metrics:
  1. *The number of external datasets accessible by the user of our prototype* measures the ability of the system to interoperate with other WP sites and to interface with the large variety of datasets.
  2. *The number of WP sites accessible by the user of our prototype* measures the compliance of the system with the federation's communication protocols and standards.
- ***Development of InfoPyramid.*** To measure the progress in developing the system for federated content-based search, we propose two metrics:
  - *The number of sectors of InfoPyramid implemented in the prototype* measures the progress in integrating multiple abstractions and modalities of the data within a unified representation.
  - *The number of types of methods implemented for abstraction and modality translation* measures the potential benefit to the federation provided by automatically substituting for missing data, reducing data resolution for transmission, and so forth.
  - *Number of software components developed.*

### 3.2 Earth System Science Metrics

We propose four distinct metrics to measure the progress of the proposed innovative approach for model development for earth sciences studies. These metrics measure innovations and new services, rather than improvements of existing products or services.

- ***Number of Implemented Functionalities.*** We identify seven distinct functionality areas: 1) Extensions to the concept of object in the content-based search framework; 2) Incorporating in the system adaptive similarity measures and iterative refinement of

models and queries; 3) Query evaluation; 4) Query formulation; 5) Visualization of query results; 6) Spatial data mining; 7) Efficient feature space indexing.

For each of the 7 areas we will identify essential, relevant functionalities and optional functionalities. The metrics starts at a value of zero. Whenever all the identified essential or relevant functionalities for one of the areas are fully integrated, the metrics will increase by 1. Thus, a value of 14 for the proposed metrics implies that all the proposed functionalities have been incorporated in the system.

- **Scalability Metrics.** To measure the scalability of our approach, we propose a metric based on retrieval speed. A benefit of integrating content-based search with model development is a sizable reduction in the data volume transferred for a particular analysis. We propose measuring a ratio of speed of analysis performed in the “naive” fashion, with all data retrieved and analyzed, to the speed of analysis with content-based search and data mining techniques employed for a set of several scenarios.
- **Model Portability.** An essential part of our research is the ability to apply local statistical models developed using local data to large scale problems. We expect that the accuracy of such models will decrease with the distance in time and location from the area used for the generation of the model, and at the same time we envision that our proposed technologies will mitigate substantially such problems. To measure our progress in this area we propose a metric based on accuracy vs. data volume curve (contingent on the availability of response data).
- **Usability .** The last metrics relates to the usability of our working prototype. For lack of universally accepted measures of usability, we propose a metrics based on a user satisfaction survey.

#### 4. MANAGEMENT APPROACH

Chung-Sheng Li will be the PL for this agreement. Lawrence Bergman, Vittorio Castelli, John Smith, and Alexander Thomasian will be PM’s from IBM T. J. Watson Research Center. Gregory Gurri Glass, Subhash Lele, Jonathan Patz, Jonathan Samet, Hugh Ellis and Robert Gillman will be PM’s from John Hopkins University, and Howard Burrows will be PM from Hughes STX. The project will also employ 3 technicians, 4 ½ graduate students, 1 payed and 2 unpaid consultants.

Lawrence Bergman will be responsible for developing the query and data mining interface for the scenarios, as well as the system integration. Vittorio Castelli will be responsible for developing data mining and knowledge discovery algorithms, John Smith will be responsible for developing storage subsystem and query planning tools, Chung-Sheng Li will be responsible for developing composite object processing and iterative refinement algorithms. Alexander Thomasian will be responsible for developing RCSVD.

Gregory Gurri Glass will be responsible for overseeing coordination of the epidemiologic analyses for the various scenarios as well as taking prime responsibility for epidemiologic interpretation of the Hantavirus and Lyme disease studies. Subash Lele will serve as the biostatistician responsible for identifying appropriate models for the spatial data in the scenarios and will cooperate with Lawrence Bergman and Vittorio

Castelli in the epidemiology-related aspects of the systems development. Jonathan Patz, the director for the JHU Program on Health Effects of Global Environmental Change, will be responsible for integrating the CAN within the program's current structure, as well as coordinating activities for the malaria scenario in India and the fire ant scenario. Jonathan Samet and Hugh Ellis will be responsible for integrating epidemiologic data associated with air quality studies. Robert Gillman will coordinate epidemiologic studies of the evolving pattern of malaria in Peru. Three full time technicians will be employed to coordinate analytical activities at JHUSHPH for the program: a system analyst, a GIS coordinator and an assistant GIS coordinator for epidemiological data.

Howard Burrows' role on will be to coordinate with other WP Federation partners and Digital Library Initiative grant recipients

Three full time students and three half time students will also participate. The full time students will work with Drs. Gurri Glass, Lele and Ellis on their specific projects. Three half-time student will be required to conduct field validation for the fire ants project. Three consultants will collaborate on the project: Menno Bouma will provide the malaria databases from India as well as other international sites and will provide guidance in use of the IBM software to predict malaria in regions undergoing landuse and/or climatic change; Dana Focks, and Rick Brenner will provide USDA data related to the application of remote sensing to the agricultural and public health analysis of Imported Fire Ants.

## 5. PERSONNEL

**Dr. Chung-Sheng Li** (100%), PL, is the Manager of the Image Information Systems Group at IBM Thomas J. Watson Research Center. His research focuses on the development of compressed-domain and progressive methods in content-based retrieval systems.

**Dr. Lawrence D. Bergman** (100%), PM, is a Research Staff Member in the Image Information Systems Group at IBM Thomas J. Watson Research Center. His research focuses on visualization and user-interface design.

**Dr. Vittorio Castelli** (100%), PM, is a Research Staff Member in the Image Information Systems Group at IBM Thomas J. Watson Research Center. His research focuses on the development of statistical methods for image analysis.

**Dr. John R. Smith** (100%), PM, is a Research Staff Member in the Image Information Systems Group at IBM Thomas J. Watson Research Center. His research focuses on image retrieval, analysis and compression systems.

**Dr. Alex Thomasian** (50%), PM, is a Research Staff Member in the Image Information Systems Group at IBM Thomas J. Watson Research Center. His research focuses on development of fast indexing methods for multidimensional data.

**Dr. Gregory Gurri Glass** (25%), PM, is an Associate Professor in the Department of Molecular Microbiology & Immunology, JHU School of Hygiene and Public Health (JHUSHPH). His current research focuses on the environmental risk assessment of various infectious diseases using remotely sensed data and geographic information systems.

**Dr. Subash Lele** (25%), PM, is an Associate Professor in the Department of Biostatistics, JHUSHPH. His current research focuses on spatial analyses and the

development of relevant statistical approaches to study the space-time interactions of epidemic spread of infectious processes incorporating environmental covariates.

**Dr. Jonathan Patz** (20%), PM, is an Assistant Scientist in the Department of Environmental Health Sciences, JHUSHPH. Dr. Patz directs the Program of Health Effects of Global Environmental Change and has been assessing the health impacts of climate change since 1993. Dr. Patz will be responsible for administration and will assist Drs. Glass and Lele in the coordination of these projects.

**Dr. Jonathan Samet** (10%), PM, is the Chair of the Department of Epidemiology, JHUSHPH. Dr. Samet is a recognized authority on environmental and occupational health. His work has focused on issues of air quality and its effects on human health in urban areas of the U.S.

**Dr. Robert Gilman** (10%), PM, is a Professor in the Department of International Health, JHU School of Hygiene and Public Health. He is stationed in Peru and is an associate of the Project for Health, Medicine and Agriculture (PRISMA), and directs the ICIDR (International Center of Infectious Disease Research) funded by NIH.

**Dr. Hugh Ellis** (20%), PM, is Professor and Chairman of the Department of Geography and Environmental Engineering of the Johns Hopkins University. Dr. Ellis will work with Dr. Samet on the public health effects of air pollution scenario.

**Dr. Howard Burrows** (25%), PM, Hughes STX is a Chief Systems Programmer. He works with NASA Digital Libraries Technology program coordinating activities between grant recipients and the science community.

**Consultants:** **Menno Bouma, MD, Ph.D** (\$35,000 /yr); **Dana Focks, Ph.D**, Senior Scientist : (In Kind -\$0-); **Rick Brenner, Ph.D**, Senior Scientist & Research Leader: (In Kind -\$0-).

**Technicians (3).**

**Students (4 1/2).**

## **6. PROPOSED COSTS**

## **7. COOPERATIVE AGREEMENT PAYMENT SCHEDULE**

- Y1, Q1:** Specify adaptive storage subsystem to provide progressive representation for storing images; Specify the query and modeling front end for model specification and validation.
- Y1, Q2:** Develop the storage subsystem; Develop the query and modeling front end for model specification; Specify the federated search engine.
- Y1, Q3:** Develop the query and modeling front end for model validation; Specify scenario 1: Hantavirus; Develop the query and modeling front end for model validation.
- Y1, Q4:** Specify the query and modeling front end for model generation; Demonstrate the generation of risk assessment maps for Hantavirus.
- Y2, Q2:** Develop the query and model front end for model generation; Specify scenario 2: Lyme disease.
- Y2, Q3:** Demonstrate the generation of risk assessment map for Lyme disease
- Y2, Q3:** Specify scenario 3: malaria in India; Specify scenario 4: malaria in Peru
- Y2, Q4:** Demonstrate scenario 3: malaria in India; Demonstrate scenario 4: malaria in Peru
- Y3, Q1:** Specify scenario 5: Fire ants
- Y3, Q2:** Demonstrate scenario 5: Fire ants; Specify scenario 6: air pollution
- Y3, Q3:** Demonstrate scenario 6: air pollution
- Y3, Q4:** Complete delivery of full system functionality to ESIP-WP federation