DISSERTATION

A GLOBAL ORGANISM DETECTION AND MONITORING SYSTEM FOR NON-NATIVE SPECIES

Submitted by

James J. Graham

Department of Forest, Rangeland, and Watershed Stewardship

In partial fulfillment of the requirements

for the degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2006

COLORADO STATE UNIVERSITY **October 24th, 2006**

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY JAMES J. GRAHAM ENTITLED "A GLOBAL ORGANISM DETECTION AND MONITORING SYSTEM FOR NON-NATIVE SPECIES" BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Advisor

Co-Advisor

Department Head/Director

ABSTRACT OF DISSERTATION

A GLOBAL ORGANISM DETECTION AND MONITORING SYSTEM FOR NON-NATIVE SPECIES

Invasive species are one of the greatest challenges to maintaining biodiversity and are a threat to human health. The Global Organism Detection and Monitoring system (GODM) has been created to allow users to: (1) manage invasive species research projects; (2) contribute data to a global invasive species dataset; (3) create maps of invasive species current and potential distributions; (4) perform analysis on datasets; and (5) extract and download data for further analysis and publication. This system can be accessed by anyone in the world with a connection to the Internet, a computer, and an Internet browser. GODM must integrate with an emerging cyberinfrastructure to take advantage of data and computing resources available on the Internet. To create GODM certain technical problems had to be solved including managing an unlimited amount of vector spatial data representing species occurrences. GODM represents a shift in how

invasive species science is done, greatly increasing the pace of science based decisions with tools that are available to a much broader audience.

James J. Graham
Department of Forestry, Rangeland and Watershed Stewardship
Colorado State University
Fort Collins, CO 80523
Fall 2006

ACKNOWLEDGEMENTS

This research was funded by a NASA grant (NRA-03-OES-03); USGS Invasive Species Program; National Biological Information Infrastructure; and Colorado Agriculture Experiment Station. I received logistical support from USGS Fort Collins Science Center and the Natural Resource Ecology Laboratory at Colorado State University. Hundreds of government and non-government users have contributed data. Over twenty students and research assistants have contributed to the testing of the system. To all I am grateful.

I would also like to thank my committee, Thomas J. Stohlgren, advisor, Melinda Laituri, co-advisor, Michael Lefsky, and Charles Anderson for agreeing to work with me and for their wisdom and continued support. Catherine Jarnevich and Cathy Stewart provided helpful comments and guidance for earlier drafts of these chapters and through their experience and kindness, gave me the insight and support I needed to complete this task. Greg Newman was the one I could always rely upon to listen and take care of all the other things I needed to do. Dave Vieglais provided valuable insights into searching the DiGIR network. Without help from Johnathon Staube and Ty Bodack and the information technology infrastructure they created at the Natural Resource and Ecology Laboratory, GODM could never have been a reality. My thanks also go to Jerry Deffenbacher who provided invaluable advice and to my life-long friend Will Carter who thought it would be cool if I had a Ph.D.

TABLE OF CONTENTS

INTRODUCTION

Non-native species cost the United States an estimated $137 billion per year in control and eradication programs and in reduced agricultural productivity (Pimentel et al. 2000). In addition, invasive species are viewed as the second greatest cause of decline in species diversity after habitat destruction (Wilcove et al. 1998). Invasive diseases impact human health and are a global problem (Mack et al. 2000), as evidenced by daily news reports of Asian bird flu, West Nile Virus, or plague. Problems caused by these invasive organisms will continue to increase in severity unless we can improve our efficiency of preventing new invasions and managing existing ones.

The Internet has fundamentally changed the way that most businesses and many individuals operate on a day to day basis. Businesses have operated at a global level for centuries and individuals have maintained long-term relationships with people in other countries for centuries. While all of this was possible in the past through the use of physical postal systems or "snail-mail," the Internet has accelerated the pace of transitions many fold. Businesses and individuals can now communicate over long distances with text, photos, sounds, and recorded or live video. In addition businesses are moving to models where entire transactions and even the businesses themselves exist only in cyberspace. While scientists have been using the Internet for communication for over a decade, the process of doing science and publishing results, including this

document, remains a time-consuming process that results in discoveries that are only accessible to a small number of individuals.

This research is about making a fundamental shift in the pace of invasive species science and making the tools of this science available to a much larger audience. Just as the Internet has allowed people to conduct businesses within "cyberspace," the Internet will allow scientists to operate within a cyberspace. This will require the construction of cyberinfrastructures that meet the needs of scientists. These new structures will allow science to tackle increasingly complex problems (Gerner 1995).

Through a unique combination of web-serving technology I will add a new model of doing invasive species science. The existing model of invasion ecology is to spend years collecting field data, analyzing results, and then publishing the results through paper-journals. This restricts the pace of introduction of results of science and the access to the field data for use by others (Magnuson and Bowser 1990). Having a global Internet-based system would allow a wide variety of individuals to contribute field data, while others may do analysis, and still others publish results. Anyone could enter the system at any stage of development breaking the traditional linear process into a non-linear process where individuals could specialize.

This change represents a shift in how we do science rather than a revolution within science; however, the increase in the speed at which science can be achieved and the greater access to data and tools to a larger group of users will cause changes within science just as it has in business and personal lives. To make this system successful we will need to insure that proper credit is given (Kuhn 1962), be as open minded as possible

(Feyerabend 1998), and insure that technology improves our lives rather then making us slaves to it (Einstein 1948).

The research presented here examines the feasibility of creating an improved system for managing invasive species. The system has a global audience as invasive species are a global problem and do not respect political boundaries and with the increase in global trade, they no longer are barred by bodies of water. The system focuses on invasive species, but since other species may be important as barriers or facilitators of invasion, all species are included. This defines the system as a Global Organism Detection and Monitoring system (GODM).

One approach to establishing the feasibility of a task is to complete the task. The first chapter provides an overview of the existing GODM system and the features it provides. The system is a functional web site providing the ability for users to upload data on organism locations, obtain maps of integrated datasets, and to download data and a large number of other features required by the system. The system is being used by individuals across the United States. The remaining issues are: (1) language translation for global usage; (2) making predictive models available on the Internet; (3) access to data; and (4) management of very large geospatial datasets. Language translation solutions are available on the Internet (BabbleFish 2006). Predictive models will be added to GODM over the next several months based on existing research at Colorado State University (Chong et al. 2001). The remainder of this dissertation concerns itself with issues of access to data and management of very large geospatial datasets.

GODM provides features that allow users to add data on invasive species locations and on environmental characteristics. However, GODM has limited resources

and will need to rely upon other web sites to collect additional data. The second chapter examines the data resources that are available through web services on the Internet and the feasibility of integrating them with new resources to create a cyberinfrastructure for invasive species management. This research builds upon related cyberinfrastructures that are already being created within the geosciences (GEON 2006) and ecology (NEON 2006), and as existing web service technologies used by research and business. Access to these resources will allow GODM to greatly expand the quantity of data available, while minimizing personnel.

The technology to manage very large raster datasets, such as those from satellite and aerial photography, is in place in a variety of web sites on the Internet (ERMapper 2006). While there are solutions for managing large raster datasets, the technology to manage very large vector datasets has yet to be developed and existing solutions are well below the capabilities of the supporting hardware. Since GODM allows users to upload data, there is no limit to the amount of vector data that will have to be managed. The third chapter provides a solution to maintaining rendering performance as this dataset increases in size.

This research demonstrates that the technology required for GODM is available and can be put in place to serve a large user base. As the system matures it will provide a new and more effective means of managing non-native species issues.

CHAPTER 1 OVERVIEW OF A GLOBAL ORGANISM DETECTION AND MONITORING SYSTEM FOR NON-NATIVE SPECIES

1.0 Abstract

Harmful invasive non-native species are a significant threat to native species and ecosystems, and the costs associated with non-native species in the United States is estimated at over $120 Billion/year. While some local or regional databases exist for some taxonomic groups, there are no effective geographic databases designed to detect and monitor all species of non-native plants, animals, and pathogens. We developed a web-based solution called the Global Organism Detection and Monitoring (GODM) system to provide real-time data from a broad spectrum of users on the distribution and abundance of non-native species, including attributes of their habitats for predictive spatial modeling of current and potential distributions. The four major subsystems of GODM provide dynamic links between the organism data, web pages, spatial data and modeling capabilities. The Core Survey Database tables for recording invasive species survey data are organized into three categories; "Where, Who & When, and What." Organisms are tracked with Taxonomic Serial Numbers from the Integrated Taxonomic Information System. A custom GIS Internet solution was required to meet the requirements of adding locations of invasive species and then immediately viewing a map of their data combined with other user's data. The geographic information system (GIS)

solution provides an unprecedented level of flexibility in database access, allowing users to display maps of invasive species distributions or abundances based on various criteria including taxonomic classification (i.e., phylum or division, order, class, family, genus, species, subspecies, and variety), a specific project, a range of dates, and a range of attributes (percent cover, age, height, sex, weight). This is a significant paradigm shift from "map servers" to true Internet-based GIS solutions. The entire system is being developed with Microsoft Corporation's Server 2003 operating system and Microsoft Corporation's Internet Information Server as a web server. We used PHP to provide high-level object-oriented programming. It has very high-performance for a scripting language and allows portability to other computer hardware and software. Custom GIS libraries were created where required for processing large datasets, accessing the operating system, and to use existing libraries in C++, R, and other languages to develop the tools to track harmful species in space and time. The GODM database and system are crucial for early detection and rapid containment of invasive species (NEON 2006).

1.1 Introduction

Largely due to global trade and transportation, harmful non-native species continue to invade (and re-invade) many areas. States, provinces, agencies, and non-government organizations are challenged to expend resources wisely to remove or contain invasive species to protect native species and ecosystem services (Mack et al. 2000). Based on global efforts to cooperate on invasive species issues (e.g., the Global Invasive Species Network) and interviews with resource managers (Crall et al. 2006) in the central and eastern United States, it was clear that a dedicated system was needed to detect, map, and model harmful invaders to help combat invasive species.

Most of the information on organism distributions available on web sites provide annually updated regional information such as the Southwest Exotic Mapping Program, Non-indigenous Aquatic Species database, and the US Department of Agriculture's PLANTS database. Many web sites also provide the ability to download existing data including the Southern Appalachian Information Node of the National Biological Information Infrastructure, and the Invasive Plant Atlas of New England. Few web sites have the sophistication of VegBank (VegBank 2006), created by the Ecological Society of America's Panel on Vegetation Classification. VegBank contains species information on over 30,000 vegetation plots located across the United States. This web site allows users to upload plot information, combine it with other plot data, and download combined results. It also allows users to browse plots by region or project and examine detailed plot data, including the percent cover of species. The GLOBE web site is an interagency program funded by the National Aeronautics and Space Administration (NASA) and the National Science Foundation (NSF) and focused on teaching students how to take weather and biological measurements. The GLOBE web site allows users to enter scientific information on-line, view maps of where information is available, and download information the user has uploaded along with other user's information. The maps are rudimentary and there is no provision for the identification of native or invasive species, and thus no link to control information or long-term monitoring capability. Finally, the DiGIR system, developed by the Biodiversity Research Center at the University of Kansas, allows users to query museum biological records through a standard interface. Museums are provided the software to place their collection online

through the DiGIR web site. DiGIR is used by many museums to provide access to their databases, but does not allow users to update the data over the Internet (DiGIR).

With emerging technologies, we are now able to build global biological databases that are accessible to multiple scientists, resource managers, policymakers, and the public for a variety of purposes (Bowker 2000). The most important factors in the success of large-scale biological information systems are how they handle the complexities arising from the precise locations and coexistence of millions of species, the relationships between species and environmental elements, and the communication and coordination required to work with various human organizations (Schnase et al. 2003). Our Global Organism Detection and Monitoring (GODM) system solves these problems by focusing on a simplified representation of biologic and abiotic information emphasizing the control of invasive species using technology that is available to users via the Internet. This advanced system was designed specifically to meet the needs of resource managers, agencies, and non-governmental organizations that manage invasive species on a daily basis. Other target users include 'citizen scientists,' researchers, decision makers, and the public.

Interviews with over 20 resource managers showed that the system should allow users to enter data on the spatial distribution of invasive species from spread sheets, from Earth Science Research Institute (ESRI) Shapefiles, from geographic positioning systems (GPS), by directly typing coordinates on-line, and by clicking on on-screen views of standard topographic maps. The most requested feature was the ability to obtain a map of data that had been entered. Local resource managers and county weed coordinators wanted this map to be printable and available to the public over the Internet. Other

important features were obtaining information on how to control and mitigate invasive species, "watch lists" of possible arriving species, and notification when new species were observed in their area of interest. The ability to download data compiled by the web site from a variety of sources and incorporation of the data into standard reports was also desired. Additional beneficial features would include maps displaying the predicted spread of invasive species and analysis tools to determine the most effective control strategies in different environments. It also became clear that the system needed to be very easy to use with minimal computer expertise and had to work over standard phone-line modems, as many of the land management offices do not have high-speed Internet access, especially in remote areas and developing countries.

In addition to the features described above, the Global Organism Detection and Monitoring system is required to allow users to manage their own projects (individual datasets) with high security, convenience, and dependability. Experts can add photos, textual information, and Internet "Links" to the 'species profiles' and other areas of the web site. Bulletin boards provide information local to an area or a specific invasive organism, while an "early warning" system will provide emails to users as new species invade. Dynamic maps allow users to examine data at the finest resolution, which may only include a single *Elaeagnus angustifolia* (Russian olive) tree along a river, to large pastures infested with *Euphorbia esula* (leafy spurge); to the global distribution of a plant genus such as *Tamarix* (salt cedar). The maps display appropriate background images including standard United States Geological Survey topographic maps and political boundaries to help users locate data. The map application integrates maps available from other Internet sites through the Web Mapping Service (WMS) protocol.

There are four major sub-systems of the Global Organism Detection and Monitoring system (GODM; Figure 1-1). First, the Internet connection allows users from virtually anywhere in the world access (www.NIISS.org). Users can use standard Internet browsers such as Internet Explorer and Netscape Navigator. Second, the Web Pages subsystem provides access to the information in the database and access to spatial data. The database maintains biological and other information, while the Spatial Data subsystem manages the large vector and raster datasets required. Finally, the Modeling subsystem will provide predictive maps and will accommodate an expanding set of modeling methods. Each of these sub-systems are described in the following sections.

1.2 Database Design

The database was designed to provide a high level of flexibility while maintaining performance. It is implemented as a relational database in Microsoft SQL Server 2000. The database design is organized into the following relational components:

- Core Survey Data
- Survey Addition
- Data Addition
- Data Tables
- Map Configuration
- Download Queries
- Taxonomic Information
- Spatial Data: Structure and locations of data
- User Information: Logins, Expertise, Project affiliations

The Core Survey Data tables for recording invasive species survey data are organized into three categories; "Where, Who & When, and What" (Figure 1-2). "Where" includes the Area surveyed and its associated Spatial Data. Spatial Data could be a point, a series of line segments for a river, or a series of polygons (delineated patches of a species or an area the organism was found in) for a physical region (Figure 1-3). "Who & When" includes Organizations, their projects, visits, and treatments. Organizations (e.g., individuals, agencies, non-government organizations) are associated with the projects (e.g. datasets). A Visit represents a survey on a specific date, at a specific location, and who completed the survey. Treatments track invasive species control efforts to determine what control efforts work best in certain environments for specified costs. "What" includes taxonomic information (Taxon Units), data on the organisms found (Organism Data) and attributes of the organisms (Attributes). Organism Data represents the occurrence of an invasive species or associated species at a visit. Organisms are tracked with Taxonomic Serial Numbers (TSN) from the Integrated Taxonomic Information System. Attributes could include height of invasive plants, abundance animals or diseases; life stage for insects; or the specifics of a treatment (e.g., the concentration of a particular herbicide or pesticide). Visits can be linked to abiotic attributes (e.g., elevation, soil type, and host species). Pathways (seed dispersal by birds, bilge water, and interstate commerce) allow the route of an organism to be tracked from its source location to its new habitat (Figure 1-2).

To cover the wide variety of ways that organizations collect information, GODM allows users to describe the configuration of their data files for addition to the database. This information is saved in the Survey Addition section so, if a user has several files that

share the same structure, they do not have to re-describe the configuration. Users can also download field collection tools, called EcoNab to quickly upload files without having to define specific file configurations. Users can store their associated metadata and citations in the Data Additions section.

Data providers or users may query the database with a flexible query engine. These data are stored in Data Tables that can be used for statistical analysis, modeling, and downloading.

GODM allows users to customize maps with data on different organisms and queries based on a variety of environmental attributes. They can add informational layers from the Spatial Data subsystem or from other web services. These customized maps are saved in the Map Configuration section.

Users can search invasive species information by scientific name, common name, and Natural Resource Conservation Service (NRCS) codes for plants. Taxonomic Information includes a relational hierarchy across taxon from the Integrated Taxonomic Information System (ITIS).

The database contains vector data (e.g., points, polygons) and the locations of files containing raster data (e.g., remote sensing outputs, maps) in the Spatial Data subsystem.

To add data to the database, or to access advanced features, users must be registered with GODM. During or after registration, users can request a higher level of security to become project managers or add information as an expert. As a project manager, they can authorize other users to add and edit data for their project. Certain data can be classified as sensitive and can only be viewed by the data provider/project

manager or by users that obtain a special clearance from the project manager. Information on users and their capabilities is maintained in the User Information section.

The remainder of the database comprises over 100 tables. These contain more detailed data on map projections used, ancillary environmental data, source contact information, level of security, taxa identification certainty, data location, sensitivity, and many other requested fields for analysis and modeling. A full representation of the database schema along with a detailed database dictionary is available at www.niiss.org.

1.3 Web Interface

The Global Organism Detection and Monitoring system is specifically designed to provide "living maps" of harmful invasive species. For example, a resource manager added new data to GODM on the location of the invasive riparian plant, *Tamarix*, and then immediately view a map of the new data integrated with existing data (Figure 1-4). Users can also zoom in to USGS Quadrangle maps or zoom out to the world, print maps, and download data. The system can be integrated with existing web sites with a custom "skin."

Our entire system is being developed with Microsoft Corporation's Server 2003 operating system and Microsoft Corporation's Internet Information Server as a web server as the tools met the requirements and were familiar to the developers. Wherever possible the programming was done in PHP. PHP is the most popular web scripting language, provides high-level object-oriented programming, has very high-performance for a scripting language and allows portability to other computer hardware and software. Custom Geographic Information System (GIS) libraries were created where required for processing large datasets, accessing the operating system, or to use existing libraries.

We designed the software using object-oriented methods to speed the addition of features and minimize maintenance of the system (Kamath et al. 1993). Quality has been assured using industry standard software testing methodology.

1.4 Spatial Data

To allow users to upload a file of data with locations of invasive species and immediately see a map of their data combined with other user's data, a custom GIS Internet solution was required. The GIS solution provides an unprecedented level of flexibility in database access, allowing users to display maps of invasive species distributions or abundances based on various criteria including taxonomic classification (i.e., phylum or division, order, class, family, genus, species, subspecies, and variety), a specific project, a range of dates, and a range of attributes (percent cover, age, height, sex, weight). This is a significant paradigm shift from "map servers" to true Internet-based GIS solutions.

Vector data such as points, multi-segment lines, and polygons are stored in the database with lines and polygons compressed into a binary large object (BLOB). Both remotely sensed and GIS-based raster datasets are compressed using Enhanced Compressed Wavelet (ECW 2006) files using libraries from ERMapper. ECW allows viewing of raster data at virtually any scale and extent with little overhead. For analysis the original data are available in uncompressed Tagged Image File Format (TIFF).

To allow the user to view the entire earth and zoom to small areas, the GIS solution needed to provide displays in various projections with minimal delay. The points, multi-segment lines and polygonal data that represent various types of locations are stored in geographic projection and in the Universal Transverse Mercator projections

for the zones in which that data overlap. Raster data layers are also provided in various projections as needed.

The custom GIS solution was written in C++ and contains a large number of open-source components. Users can upload and download data in a variety of projections and datums. Proj4 (Evenden 1990) is used to project this data into Geographic and UTM projections in World Geodetic System 1984 (WGS84). The Geospatial Data Abstraction Library (GDAL 2006) provides the ability to read and write projection files. Various open-source file translation libraries are used to read and write data files. Overall GODM contains over 100,000 lines of custom PHP software and over 100,000 lines of custom C++ software.

Raster datasets such as the United States Geological Survey's 1:24,000 scale topographic maps can contain over 10,000 files in a single UTM zone. To maintain performance requirements, grids were overlaid on these file sets. The database stores the rows and cells of spatial data from the grids, allowing database indexing to be used to find the files that overlap with a specified viewing area and restore search times (Longley et al. 2001).

1.5 Integration of Modeling

The GODM system provides an extensible architecture to support many existing modeling methods (e.g., Morisette et al. 2006) and add new methods in the future (Figure 1-5). Complex spatial models, such as the NASA Invasive Species Forecasting System, which use multiple remote sensing layers over large spatial extents, and high-performance computing, will be run as a separate process on one or more computers.

Limited spatial models (e.g., for small areas or simple Kriging models) are being developed in GODM using the open source statistical package, "R", which is integrated into the system to provide a wide range of descriptive statistical calculations. Custom software was added to allow control of processes including the setting of priorities, collection of output, and the ability to terminate jobs as these feature were not available in the standard PHP libraries. Custom components provide high-performance autocorrelation checks and the generation of regression surfaces. The open source Geospatial Statistical Library (GSLIB) provides semi-variograms and Kriged surfaces.

In the future, the more complex multivariate and time-space modeling features made available through NASA's Invasive Species Forecasting System will be more closely exchanged with species data from the GODM system to bring a far greater array of high-performance predictive spatial modeling tools to our variety of user communities.

1.6 Hardware Architecture

The hardware for the GODM system has been designed to provide visualizations and analysis on large geospatial data sets very quickly (Figure 1-6). Requests are entered through the Internet and are routed to a load balancer on the main system or a mirrored system. The Load Balancer sends requests to an idle Web Server to insure users have a quick response. Long processing jobs are submitted to the database and are picked up by Compute Servers. The GIS Servers provide high-speed access to large raster datasets.

To insure performance and reliability the entire system is mirrored at multiple locations. The web servers are housed at the United States Geological Survey's Fort Collins Science Center and the Natural and Environmental Sciences Building at Colorado State University. The servers will communicate with the National Aeronautics and Space

Administration (NASA) high-performance computers associated with the Invasive Species Forecasting System and other organization computers through standard web-services protocols.

1.7 General Application across the Globe

The GODM system represents a significant shift in the way invasive species information is managed and shared. GODM will allow academics, resource mangers and the public to share data and participate in group analysis and decision-making about the way we manage biological resources, and especially harmful invasive species. The technology is applicable to other areas of natural resource management including the management of threatened and endangered species, fire management, and for tracking wildlife or human diseases.

Future features will include: additional modeling approaches, new statistical methods, additional GIS layers, additional remote sensing layers for more accurate predictive modeling, and a protocol to connect GODM to other databases that contain biological data. Over the next two years, additional modeling strategies for controlling invasive species also will be added. At the time of publication the system includes data from 69 projects, including 37,495 field surveys, with over 130,000 organisms identified from 1562 different taxa.

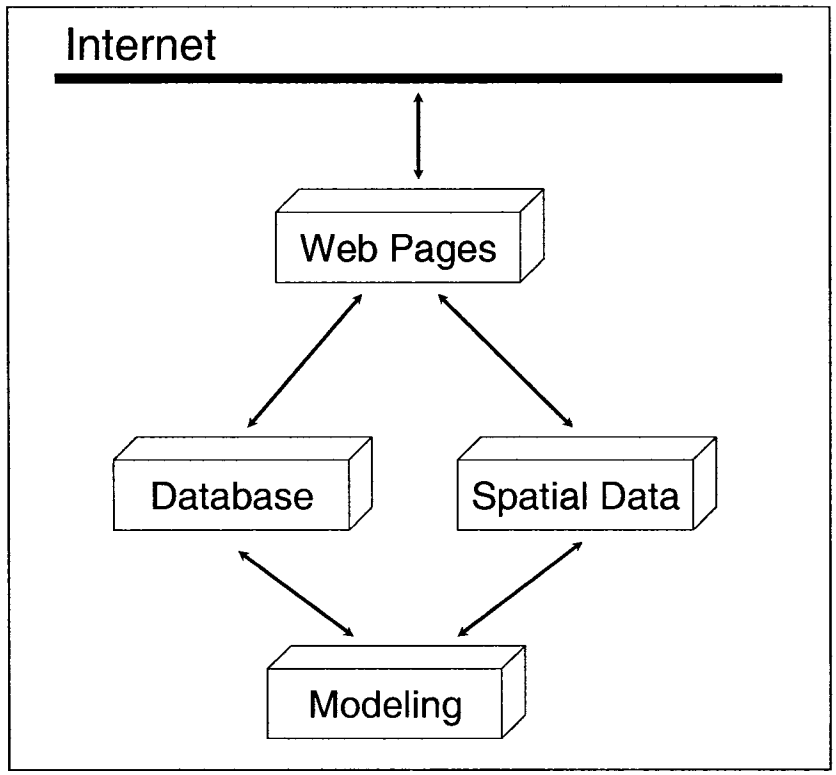

Figure 1-1. The four major subsystems of the Global Organism Detection and Monitoring system (GODM). Web Pages provide dynamic access to the database and available spatial data. Modeling provides predictive maps.

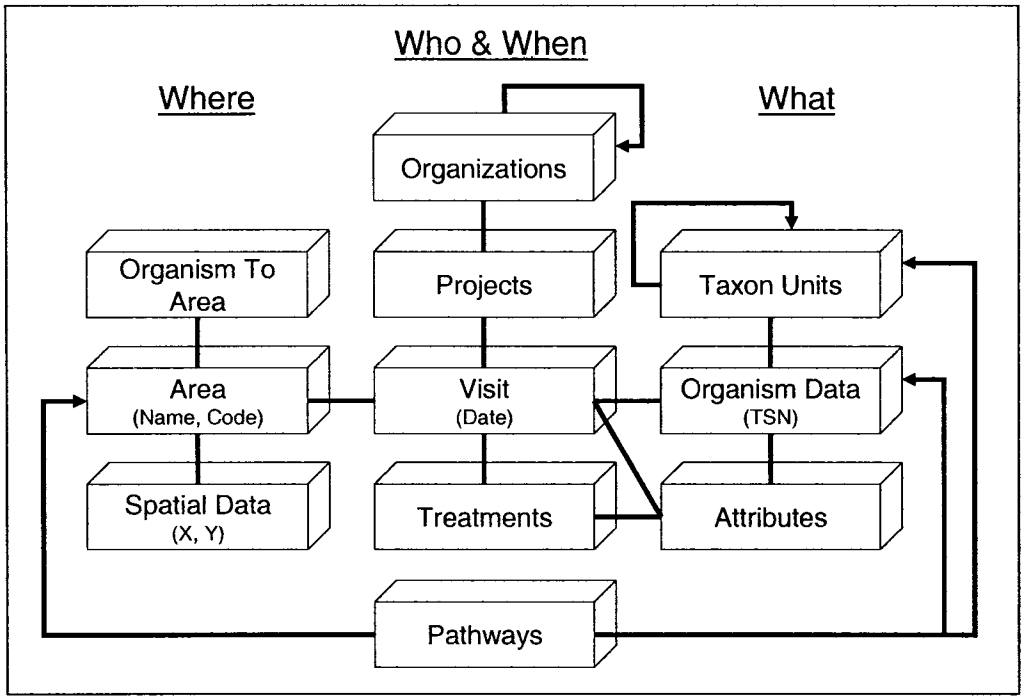

Figure 1-2. Database tables for organism survey data.

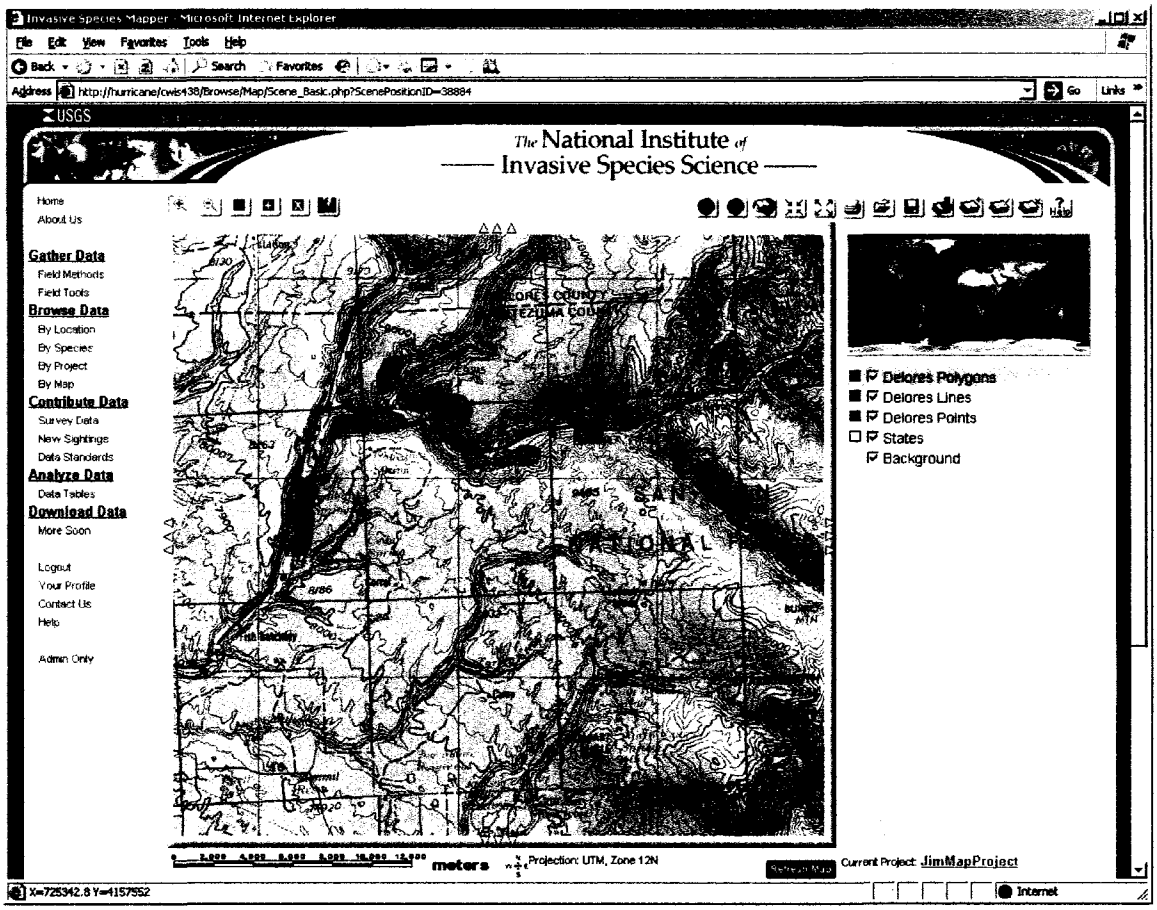

Figure 1-3. Simulated map showing the three types of spatial data supported for surveys: points, polygons, and multi-segment lines

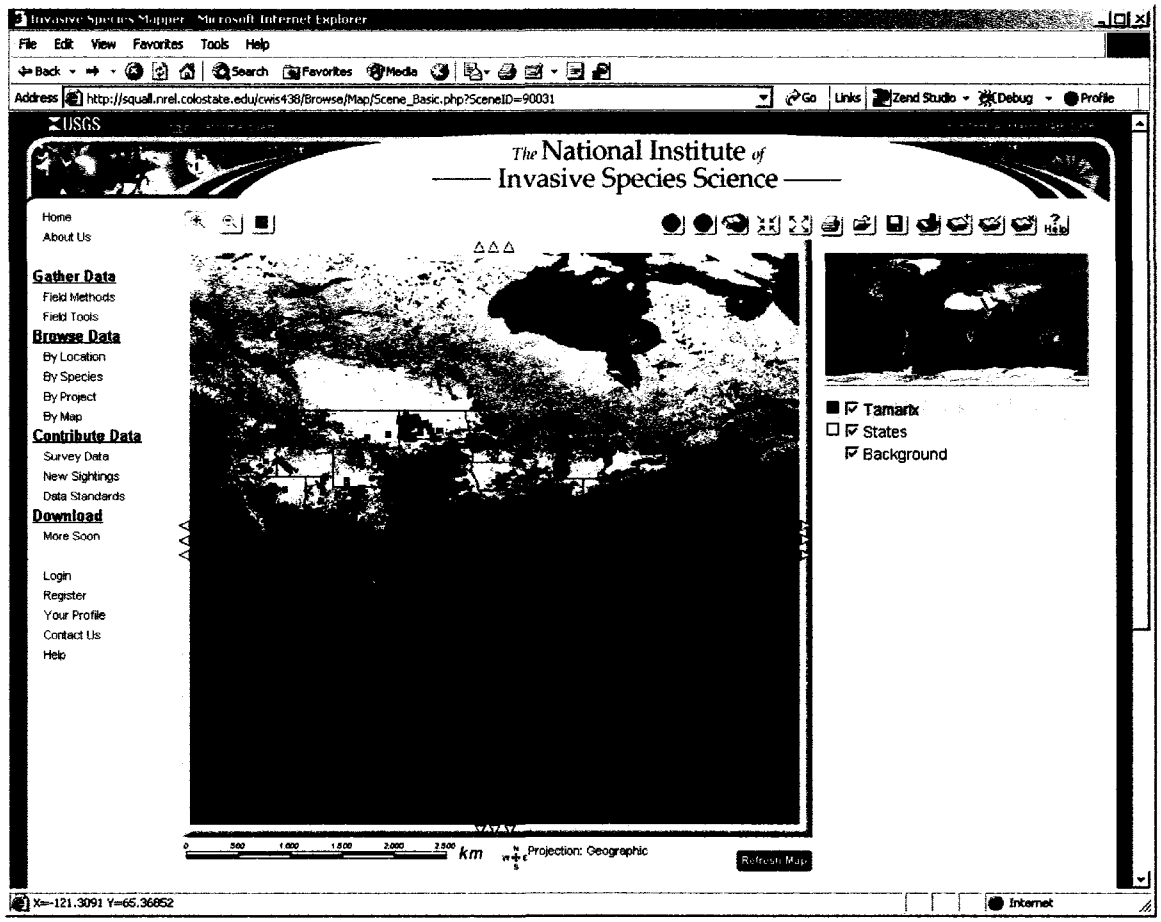

Figure 1-4. The map viewer page showing the actual current data on the distribution of *Tamarix* in the United States from over 100 independent databases uploaded into the GODM database.

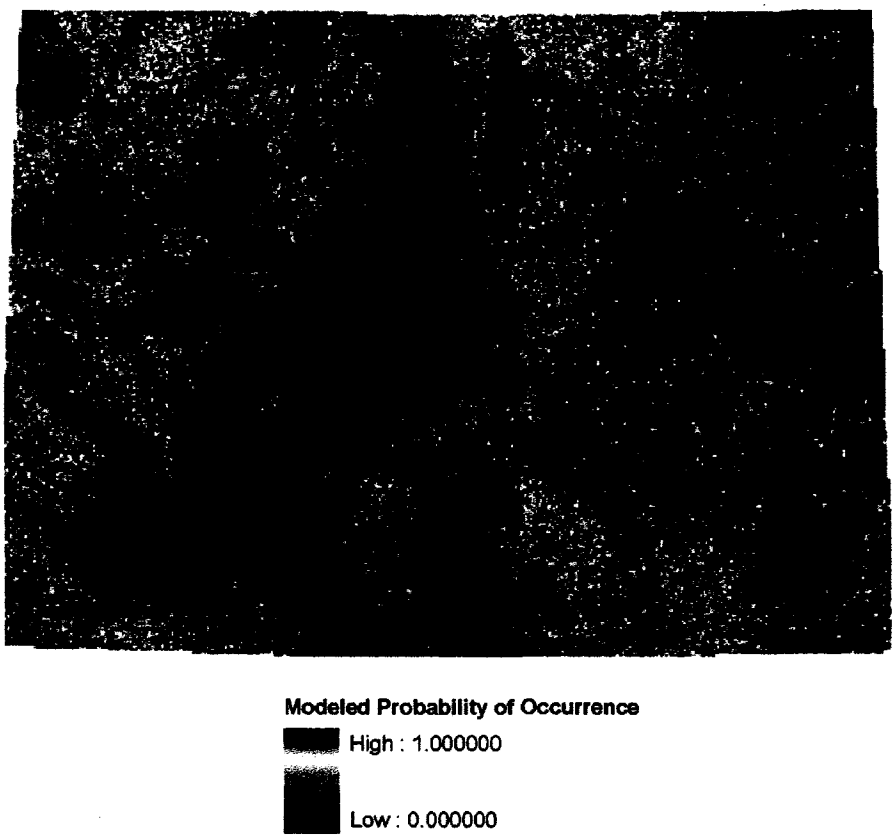

Figure 1-5.Predicted probability of occurrence of *E. esula* in Colorado from the general linear model and Kriged residuals (Crosier 2004).

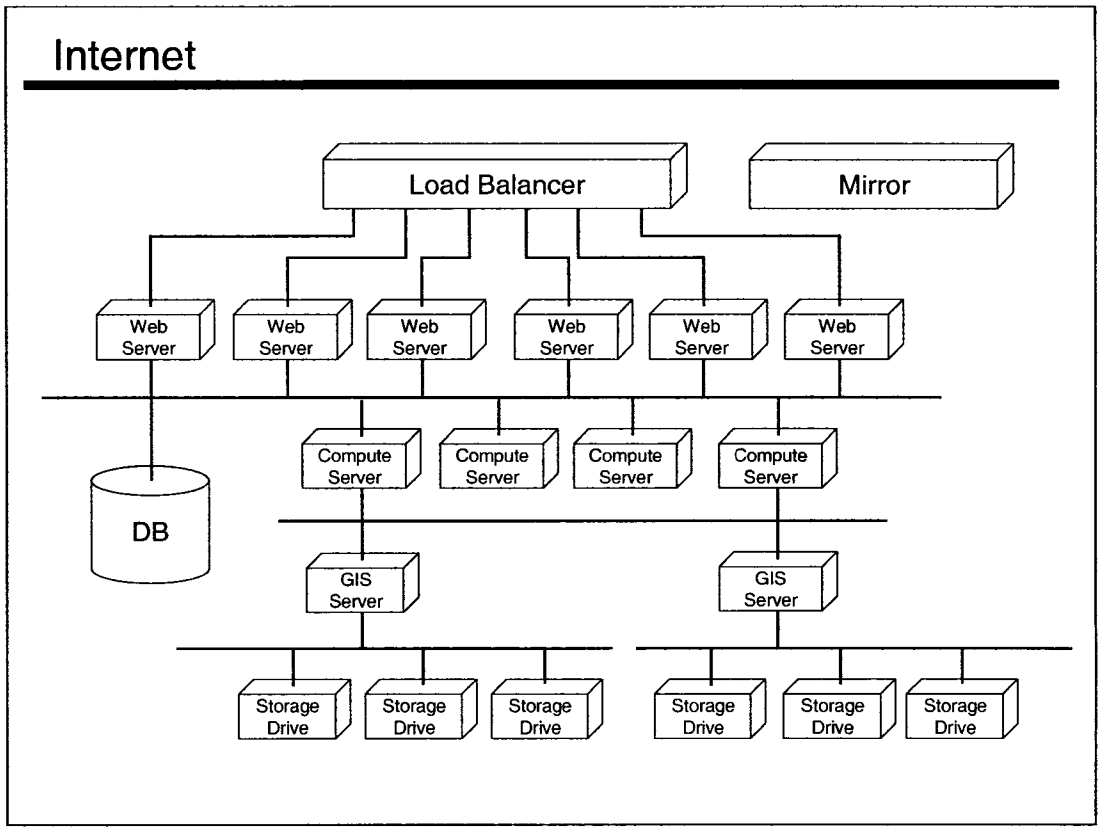

Figure 1-6. Hardware components for the Global Organism Detection and Monitoring system.

CHAPTER 2 CYBERINFRASTRUCTURE FOR DOCUMENTING, MAPPING, AND MODELING NON-NATIVE SPECIES ABUNDANCE AND DISTRIBUTION WORLD-WIDE

2.0 Abstract

Real time, accurate data on the location and abundance of harmful non-native invasive species are needed to contain them at local, regional, national, and global scales. The strategy presented for a comprehensive Invasive Species Cyberinfrastructure is based on the commonly accepted strategy for managing invasive species including: (1) watch lists and early detection of new invaders; (2) modeling the current and predicted extent of an invasive species range and abundance; and (3) applying "best management practices" for control and restoration efforts. Modeling ranges and abundance requires access to species occurrence data and environmental data. While the exchange of environmental data over the Internet in the form of raster data is maturing, the exchange of species occurrence data is just emerging and it can take years to access the available data. These performance problems can be addressed by improving the web service protocols, the server software, and the database queries used to access the data to significantly reduce the time to harvest data.

2.1 Introduction

The spread of harmful non-native species is growing as globalization of commerce has increased the movement of terrestrial and aquatic organisms (Mack et al. 2000, Stohlgren et al. 2006). Consequently, our inability to efficiently and effectively combat these invasions has resulted in enormous environmental and economic losses worldwide (Pimentel et al. 2000).

Documenting, mapping, and modeling harmful invasive species requires detailed geospatial data on species abundance and distribution relative to environmental characteristics of favorable and unfavorable habitats (Graham 2006, Stohlgren and Schnase 2006), in addition to information on cost-efficient site-specific control techniques. Finding and containing invaders early, while populations are small and concentrated, is the first step in an "early detection and rapid response" program because small populations can be feasibly controlled (Rejmánek and Pitcairn 2002). This requires field personnel to be equipped with effective tools to identify and record invasive species and habitat characteristics, and immediately report them to concerned individuals and agencies. Immediate information sharing to surrounding people and agencies would enhance comprehensive programs in the prevention, early detection, containment, and monitoring of harmful invaders at local, regional, and national scales (Stohlgren and Schnase 2006).

In the past, ecologists have devised and executed field experiments, analyzed their results and published them in journal articles. Analyses may have included basic statistics, evaluating species-environment relationships, and mapping species distributions post-hoc. Also, time between fieldwork and publication could be years.

I propose a paradigm shift where ecologists place their data online for others to access, request data from online sources from other contributors, perform analyses online, and immediately publish their results online. Such a system would free scientists and resource managers from having to follow a linear process. Instead they could take existing data and analyze it or publish an article based on multiple versions of analysis that were already completed in a shortened timeframe. This system is a cyberinfrastructure for ecologists and similar systems are emerging in other fields such as geology (GEON 2006). The system would integrate data requirements for analysis and mapping, with field technologies (e.g., palm computers, global positioning systems), to ensure the rapid assimilation of accurate field data along with environmental data derived from satellite imagery and GIS. The cyberinfrastructure will include high-performance computers, large storage computer servers, and a network to connect them together as described in more detail below.

2.1.1 Challenges to Mapping and Forecasting

Determining where to survey for existing invasions is difficult for all but the most obvious and generalist invaders, such as kudzu or cheatgrass. Habitats vulnerable to invasion may be patchily distributed relative to source populations. The species may be cryptic and difficult to detect, hiding in the seed bank or over-wintering in another habitat. Metapopulation dynamics may produce small populations in suboptimal habitats for later dispersal to optimal habitats, confounding our understanding of species-environment relationships. Likewise, uncertainty about dispersal, establishment, growth, and survivorship in various habitats can greatly affect the design of effective surveys (Stohlgren 2006).

Additional questions remain on how best to control invasive species. Different combinations of manual, chemical, and biological control might be more effective in different habitats. These in turn, may be affected by the proximity and size of source populations. These questions involve complex characteristics of invasive species dynamics and their interactions with environmental characteristics and other species (Stohlgren and Schnase 2006).

Thus, mapping and forecasting species abundances in space and time is a complex statistical modeling process (Stohlgren et al. 2006). The integration of ecology with mathematics and statistics has provided a solid analytic foundation for these models, but with additional complexity. Mapping and modeling are required at small scales to aid resource managers in making specific decisions on how to manage individual infestations and also across continental scales to aid decision makers in the allocation of resources for prevention and screening (Green et al. 2005).

2.1.2 Strategy

The approach for a comprehensive Invasive Species Cyberinfrastructure is based on the commonly accepted strategy for managing invasive species including: (1) watch lists and early detection of new invaders (Drucker 2007); (2) modeling the current and predicted extent of an invasive species range and abundance (Stohlgren and Schnase 2006); and (3) applying "best management practices" for control and restoration efforts (TNC 2006).

Developing "watch lists" for invasive species requires an understanding of invasive species attributes adjacent to the target area, the vulnerability of habitats to specific invaders inside the target area, and the potential for spread from outside to inside the target area (Drucker 2007). All these are multivariate problems with complex

modeling solutions. Early detection involves "smart surveys" (Stohlgren and Schnase 2006) of the most vulnerable habitats and the most invasive species to detect populations while they are small and can be affordably controlled or eradicated.

Modeling the current and potential range and abundance of an invasive species can be done with information on its existing and native ranges and abundances, and the vulnerability of habitat to invasion. This task requires field personnel across the earth to collect information on the location and abundance of the potentially invasive species. Data collection can be made more effective by examining the realized niche (or environmental envelope) of invaders relative to potential habitat for invasion (Barnett et al. 2006). Monitoring the location and abundance of an invasive species again relies heavily on field personnel and a stream of real-time data.

The best management practices for control and restoration efforts for an invasive species can vary based on the climate and the resources for treatment that are locally available. For this strategy to be viable there needs to be an efficient system that allows information exchange between inspectors, field personnel, modelers, and resource managers to better coordinate activities from local to global levels.

2.1.3 Invasive Species Cyberinfrastructure Requirements

There are varying requirements for documenting, mapping, and modeling non-native species. Documenting non-native species includes their taxonomic identification, characteristics (or profile), and occurrence in space and time. In addition, metadata could be added to this information to allow identification of when the data was collected, who collected to the data, and how the data was processed. A more comprehensive list of

potential metadata is contained in the Federal Geographic Data Commission (FGDC) standard.

The minimum requirement for mapping a species is a location on the earth and a taxonomic identification. For users of the cyberinfrastructure to create maps with the locations of a given species based on temporal information then a date of collection must be included. The minimum requirements for modeling species distributions are a taxonomic identification, whether the species was present or absent, a location on the earth, and to create dynamic models at date when the species was observed. Modeling species abundance requires some attribute describing abundance of the species at each location. A wide range of additional attributes, environmental values, and control data could also be used in modeling. The combination of a taxonomic identification (typically a species), a location on the earth, and a date of observation are all common to documenting, mapping and modeling non-native species.

Currently it is very difficult to obtain a large collection of data with the locations of non-native species (Crosier 2004). Data on the location of invasive species is collected through observations or field surveys, which is collected either on paper or electronically. Some of this information is then available in databases (Crall et al. 2006). The Global Invasive Species Information Network (GISIN) maintains a list of over 200 online databases with a large variety of documentation on invasive species (GISIN 2006). Web services provide the potential for other computers to access these and other databases over the Internet and there is currently a GISIN proposal for an Invasive Species Profile Schema (GISIN 2006) to facilitate information exchange.

The Global Biodiversity Information Facility (GBIF) states that there are approximately 1.5 billion specimens in herbaria and museum collections worldwide (GBIF 2006). The Distributed Generic Information Retrieval (DiGIR) is a web service protocol that uses the DarwinCore schema to exchange data between computers, referred to as servers. GBIF currently indicates there over 600 online databases containing over 66 million specimen records. DiGIR provides a potential source of information on organism occurrences and a test of feasibility for the exchange of organism occurrence data across a biological cyberinfrastructure. However, museum records may be patchily distributed or out-dated for fast moving invasive species.

There is a wide range of environmental information available for modeling non-native species that could be used for modeling invasive species abundance and distributions. This data includes Web Mapping Service (WMS 2006) sites such as the National Aeronautics and Space Administration's (NASA) Jet Propulsion Laboratory (JPL 2006) web site which provides remotely sensed data including digital elevation models. The WMS protocol allows web servers to call the JPL site and request portions of remotely sensed data in various resolutions and for various areas. Systems such as the National Digital Forecast Database (NDFD 2006) use web services to provide weather information for locations in the United States. Systems such as the National Ecological Observatory Network may provide additional web services to access environmental information (NEON 2006) in the future.

The objective of this research is to determine if a cyberinfrastructure is feasible to aid the documenting, mapping, and modeling of non-native species current distributions, predicted distributions, and control strategies with acceptable performance to end-users.

This system could provide at least national, and potentially global, statistical predictions of non-native species existing ranges, future ranges, and effective control strategies. This cyberinfrastructure will need to integrate: (1) the latest information non-native species location data; (2) temporal data of species locations; (3) organism attributes; and (4) ancillary data on environmental characteristics where non-native species are found (or not found). This cyberinfrastructure will also need to provide access to the data through different user interfaces (web-sites).

Data sources include paper-based systems, computer files, isolated databases, web sites, and web services. This research focused on the availability of occurrence data and environmental characteristic data through web services. Since web services are an existing technology the primary concern is the potential performance and reliability of these services.

2.2 Methods

A cyberinfrastructure contains a set of computers, referred to as servers, containing databases and other software, that communicate through the Internet using web service protocols. Each of these components can affect the overall performance of the system. The Internet is a complex system and performance problems can occur based on the location of servers, the amount of traffic (which changes during the course of a day), the type of server hardware, and web server software. With online databases the design and number of records in the database can also affect performance. For web services the complexity of the protocol and the type and design of the web service software can significantly affect performance. Testing each of these components individually is not feasible within the scope of this project. Instead, this study will

examine the performance characteristics of existing web services that use two protocols, DiGIR and WMS.

2.2.1 DiGIR

The DiGIR network was created to exchange specimen data from museum and herbaria over the Internet. This system contains over 100 servers across the earth and can provide a test environment to evaluate the feasibility and issues associated with biological web services. DiGIR is a client-server protocol but refers to the client as a "portal" and the server as a "provider." Providers register as "businesses" with the Universal Description, Discovery and Integration (UDDI 2006) protocol. Businesses are registered with the services they provide and a uniform resource locator (URL), or "AccessPoint" in DiGIR terminology, which is used to locate the provider server on the Internet. The metadata for each service/URL can be requested from the URL and contains the "codes" for each "resource" at the service. Each resource is effectively equivalent to a database that contains data on biological specimens or observations.

The DiGIR protocol allows resources to be inventoried and searched from other servers on the Internet. Inventory refers to querying the number of records of a certain type, such as querying the number of genus and receiving a list of each genus name with the number of available records for each genus. Searching refers to providing a Boolean search string for a given field in the database and then receiving some number of records that meet the search criteria. These records include DarwinCore 1.0 fields such as ScientificName, YearCollected, and Latitude (DiGIR).

DiGIR was evaluated by building web crawlers that harvest data from the existing DiGIR network of providers (Fox et al. 2004). Measurements taken during harvesting

included performance, robustness, and numbers of records obtained. The results of these tests were used to determine if it is feasible to obtain data from web services and the characteristics of such a system.

The performance measures determine the feasibility of creating maps on the fly using data directly from DiGIR providers. Data acquisition is expected to take too long to produce a map in the time required by users but the information gathered will also be used to determine if it is possible to harvest data from DiGIR providers and store it locally on a monthly basis. The Global Organism Detection and Monitoring (GODM) system has already shown that it is possible to render maps quickly with local data (Graham 2006). The data will also be used to infer if it is possible to harvest biological occurrence data from the Internet in general and could impact the definition of proposals for invasive species data protocols and biological protocols in general.

2.1.1 Harvesting DiGIR Business, Service, and Resource Information

The first step was to create a system to acquire the Businesses, Services, and Resources available from the GBIF UDDI database and each DiGIR provider. I then stored this information locally for future searches. In an operational system, this information will have to be updated on a regular basis. Then, it was measured for performance and quality. The performance measures included the time required to obtain business, service, and resource information.

Once the data were acquired, I determined the number of services per business and the number of resources available per service. The resources also provide information on the number of records in each resource, which I used to analyze the performance of the searches.

Businesses were searched alphabetically by submitting the search string "A%", then "B%", etc. where the percent is a wild card character. The information for each business contains a BusinessKey, which was used to query the service information including the AccessPoint and a ServiceKey. I used these values to request the metadata from each provider. The metadata contain the ResourceCodes for each resource within the service and information on the number of records the resource contains. All this information was stored in a local enterprise level relational database for analysis and to provide information to harvest records from DiGIR providers.

Since the businesses were harvested in blocks based on the first letter in the business, a breakdown of performance by business was possible. Service details were acquired for each business so I could analyze the time to acquire this information on a business basis giving the total, mean, minimum, and maximum times. The time required to request and receive the metadata for each service was analyzed in a similar manner. I then evaluated an equation for the expected performance given a projected number of resources was then created. This equation determined the expected performance as the number of resources increases over time.

2.2.1.2 Harvesting DarwinCore Records

I built a separate crawler to search for data on a specific genus or a scientific name (genus and species combined) across all available DiGIR resources. This crawler was used to complete eight harvesting operations; *Tamarix sp.* (Tamarisk), *Bromus tectorum* (Cheatgrass), *Solenopsis invicta* (Fire Ant), *Myocastor coypus (*Nutria), *Sturnus vulgaris* (European Starling), *Pueraria montana* (Kudzu), *Dreissena polymorpha* (Zebra Mussel), and *Lymantria dispar* (Gypsy Moth) across all available DiGIR providers. The

searches were made using the "LIKE" comparison operator against the scientific name with the "%" wildcard at the end of the search string to include all species with a genus and all varieties and subspecies within a species. Performance measures were made as the search was executed on each provider and included; the number of seconds for the request to be completed, the number of seconds to parse the response, and the number of matches). Parsing accounted for less than ten percent of the search time and was therefore not included in the analysis. The results of the searches were stored in a local database for analysis.

There are two methods to specify the fields in a DiGIR search request, directly or by the use of an extensible markup language (XML) schema definition (XSD) file on another server. The examples provided with the DiGIR software show using an xsd, but this was found to slow the searches and periodically failed due to a denial of service at the SourceForge web site that contains the xsd files. Instead, all fields were specified directly and only the most critical fields for modeling were queried, including DateLastModified, BasisOfRecord, InstitutionCode, CollectionCode, CatalogNumber, Latitude, Longitude, Phylum, ScientificName, YearCollected, MonthCollected, DayCollected, and TimeCollected.

The number of records that can be requested in one search operation varied by resource, with a minimum of 100 and a maximum of 50,000 records. To make a systematic comparison of the resources the value of 100 records was used for all resources.

The response durations were compared with the number of records in each resource and with the number of matches to see what the driving variables are in the

DiGIR system. The top 10 providers for the most successful harvesting operation, *S. vulgaris*, were examined for a systematic change in harvest time based on number of records contained in the provider, number of matches for a search, and by search duration.

Early testing showed that there was a significant performance loss based on the number of records per request. To understand this effect, three tests were performed, one with only a database query, one with a "Simple DarwinCore Request" and one to the existing DiGIR Provider solution. These tests were with a range of numbers of records per request and recorded the time required to obtain all the data for one species. All tests were run on a local computer at Colorado State University using data harvested from DiGIR providers, specifically from the request for the *S. vulgaris* occurrences. The three tests used the same query on the same server, database and data. This removed variation based on the Internet connection, hardware, and web server and database software to allow a comparison of just the differences between request methods. Times to harvest the data were recorded as a function of the number of records received with a constant request size of 1,000 records per request and by holding the number of records constant and varying the number of records per request.

I used the structure of a DiGIR request to develop the Simple DarwinCore Request. There are effectively two critical features of a DiGIR request: (1) specifying the fields to be returned; and (2) specifying the Boolean search to be completed. In almost all cases, the databases of DiGIR providers will be based on the SQL language. This implies that the fields specified in the DiGIR protocol are translated into fields in the provider's database and placed in the "SELECT" portion of a SQL query. The Boolean

search would then be translated and placed in the "WHERE" portion of the query (Celko 2005). The optimal method for encoding these elements would appear to be a comma-separate list for the fields and a traditional Boolean search string for the search. Within a traditional HTTP GET/POST, the request would appear as:

Fields=Field1,Field2&Where=Species LIKE Tamarix

The SQL statement could then appear as:

SELECT Field1, Field2

FROM Table1

WHERE Species LIKE Tamarix

A specific database implementation may need to add additional tables and WHERE conditions to match the individual database design.

The DiGIR response contains a series of records. Each record contains the requested fields with its name and its value. For the Simple DarwinCore Request, this information was returned with a simple Extensible Markup Language (XML) two-level hierarchy with the first level being each record and the second level being a field with a name and value as follows:

```
<record>
    <field name='ScientificName' value='Tamarix'/>
    <field name='Latitude' value='40.0'/>
    <field name='Longitude' value='-105'/>
</record>
<record>
    ...
```

</record>

Results from the three local tests included a comparison of the request duration over the number of records and the request time over the number of records per request. Polynomial curves were fitted to these datasets to determine the characteristics of the performance under each of the three request methods.

2.2.1.3 Quality of DarwinCore Records

To use data from DiGIR within GODM, we must obtain a taxonomic identification as a Taxonomic Serial Number (TSN) from the Integrated Taxonomic Information System (ITIS). Any scientific names that cannot be matched to a current TSN were eliminated. GODM also requires at least a year for the collection date and a physical location. At this time DiGIR supports Latitude and Longitude or a textural description for physical location so all data without Latitude and Longitude were eliminated. The remaining records provided the number of records usable by GODM for a given search. To illustrate the contribution made by the DiGIR data, maps were created of the existing *Tamarix sp.* data in GODM and the data obtained from the DiGIR providers for the genus *Tamarix sp.*

2.2.2 Environmental data

The sources for environmental characteristics parallel those for organism occurrence data and included paper-based, computer files, isolated databases, web sites, web services. These data can be broadly broken down between data from remotely sensed raster data such as aerial photos and satellites, and data acquired directly from

environmental instruments such as weather stations. A web crawler was built to evaluate the reliability and performance of remotely sensed data from JPL.

The selected dataset was the National Elevation Dataset (NED) from the JPL site. The crawler requested a number of rasters from the JPL site based on predefined or randomized values and saved them locally as Geo-referenced TIFF (GeoTIFF) files. All files were requested in the Geographic projection and contain 16-bit signed integer values for elevation. Parameters included the zoom level in pixels per map unit and the area selected. The time to acquire each image was recorded. Previous tests showed that if the server failed it returned an error and/or an empty file. Both error messages and the file sizes were also recorded.

The first test made 1,000 requests for rasters with random zoom levels between 108 pixels per degree (approximately 1 kilometer per pixel) and 10,800 pixels per degree (approximately 1 meter per pixel). The area sampled was randomly selected to fit within the continental United States between latitude 25 and 48 North and between longitude –120 and –70 West. The size of the raster was fixed at 500 by 500 pixels. The second test made another 1,000 requests but added randomizing the size of the raster between 10 and 1,000 pixels.

I determined if this web service changed performance based on the amount of sampling that was required to create the requested image by analyzing performance for various resolutions. For the second test, I analyzed the request time versus the size of the rasters requested. The number of images that failed to be saved to a file was recorded and a random visual sampling of the contents of the files was completed to insure appropriate raster data were transferred.

2.2.3 Computer System

The computer system used for all testing was an IBM-PC compatible computer with 4 processors running at 2.99 GHz and 1 Gigabyte of RAM. The computer used the Windows 2003 server operating system with Microsoft's Internet Information Server (IIS) version 6.0 as the web server and PHP 5.0.5 for the web scripts. A simplified custom XML parsing class parsed responses from the servers. The computer was installed at Colorado State University's Natural Resource and Ecology Laboratory (NREL) and was connected to other servers on the Internet at speeds between 45 and 165 Megabits per second.

2.3. Results

2.3.1 DiGIR

2.3.1.1 Harvesting DiGIR Business, Service, and Resource Information

It took 21.43 seconds to harvest the information for 192 businesses from the GBIF UDDI database. Total time to harvest and save this data into the database was 132 seconds. Information for 191 services was harvested including details about the services (Table 2-1). Of these, 135 of the services could be identified as containing DiGIR resources as opposed to other protocols, and 105 of these contained one or more resources for a total of 663 resources (Table 2-2). The total number of available records for organism occurrences from these 663 resources was over 66 million.

2.3.1.2 Harvesting DarwinCore Records

The time to request data from DiGIR providers varied from 4.2 seconds to 78 seconds per record. Most of the times were between 4 and 10 seconds while the 77-second value was for *S. vulgaris* records (Table 2-3). The total time to harvest the *S. vulgaris* data was over 13 hours. The number of records explained less than 13% of the variation in search time (Figure 2-1). Request durations explained less than 3% of the variation in request time with the exception of the *S. vulgaris* search where it explained just over 17% (Figure 2-2).

When searching for data within the species *S. vulgaris*, the slowest search was from the Swedish Museum of Natural History (NRM) at 43,788 seconds or over 12 hours. This resource contains over 6,000,000 records but the largest database had over 10,000,000 records and responded in 3.52 seconds (Table 2-4). The search duration did vary systematically with the number of matches when the results were sorted by number of matches (Table 2-5) or search duration (Table 2-6). The only systematic variation found with the request duration when there was a very large number of hits (Figure 2-2.E). The two extreme points in this graph represent 8,300 records from the Avian Knowledge Network and 46,740 records from The Swedish Museum of Natural History (NRM).

Varying the number of records requested, in blocks of 1,000 records per block, showed an exponential relationship between all these methods for requesting data (Figure 2-3). Factors for the exponential component varied from 1.0 for the Simple DarwinCore Request to 9.828 for the DiGIR Provider method (Table 2-7). In all cases the polynomial, which was fitted through zero for the intercept, explained over 99% of the

variation. A zero intercept could be used because a search that resulted in zero records to a local server should be near zero (Table 2-1). Request time decreased rapidly as the number of records per block changed from 100 to 1,000 (Figure 2-4).

2.3.1.3 Quality of Records

The mean number of records usable by GODM per search was 9,568, but this was dominated by the search for *S. vulgaris* (Table 2-8). The rejection rate for records was 17.7%. Some of the providers returned the Kingdom, Family, Genus, and BasisOfRecord fields even when not requested. Other fields such as Class and Order were returned empty when not requested. The *M. coypus* data came primarily from the EUNIS resource at the European Environment Agency and was found that most of the ScientificName, Latitude, and Longitude fields were not returned from the provider.

While the data for *Tamarix* showed 641 records only 292 of these were for species that are considered non-native to the United States including 45 records for *Tamarix aphylla*, 138 for *Tamarix ramosissima*, and 109 for *Tamarix chinensis* (Figure 2-5). While this is a small number of records compared to the number of *Tamarix* records currently in the GODM database, these records represent areas not covered by the current GODM data (Figure 2-6).

2.3.2 Environmental data

2.3.2.1 Performance

The first test examining raster acquisition performance showed that the time to request rasters from JPL was not dependent on the zoom level. The times typically

varied between 1.5 and 3 seconds with a number of times near 0 for failed transfers and a few transfers taking up to almost 7 seconds (Figure 2-7).

The number of pixels requested explained most of the variation in the amount of time to obtain the raster, R^2=0.66 (Figure 2-8). Since I was not interested in the failed transfers as a component of the transfer speed, I removed the failed transfers and obtained a relationship that showed 80% of the variation in time explained by raster size, R^2=0.81 (Figure 2-9). In both tests, the y-intercept was near 1 second.

2.3.2.2 Quality

In the first test, 90 of the 1,000 rasters failed to transfer, a 9.0% failure rate. On further inspection this there were four blocks of contiguous transfers that failed with the error "Service denied due to system overload. Please try again later." A random inspection of 20 of the files found six to be a constant gray value instead of elevation values. This is to be expected when sampling outside the available data for the contiguous United States. In the second test, 55 of the 1,000 rasters failed to transfer, a 5.5% failure rate. Random inspection of 20 of the files showed that nine contained a constant gray value.

2.4. Discussion

2.4.1 DiGIR

2.4.1.1 Harvesting DiGIR Business, Service, and Resource Information

The time to harvest the GBIF businesses from UDDI was under 2 minutes. The mean time to harvest service details for each service was 1.051 seconds giving us a performance equation of:

Equation 1: $TSD = NS * 1.051$

TSD is the time to obtain all service details and *NS* is the number of services. Equation 1 shows that even with 10,000 services it will only take approximately 3 hours to harvest all the service details, which can easily be accomplished on a monthly basis.

The mean time to request and parse the resource information was 25.02 seconds giving an equation for the time to harvest resource information, *TR*.

Equation 2: $TR = NS * 25.02$

If the number of services rises to 10,000 is will take over 69 hours or almost 3 days to harvest this information. This was surprising since the metadata containing the resource information could be maintained in a static XML file. On examination it was found that this file was being created on each request and a "COUNT(*)" SQL command was being executed to determine the current number of records in the database. This is a very time consuming query, as it needs to find all records that match the search. Removal of this entry in the metadata may solve the problem and bring the time down to about 1 second.

2.4.1.2 Harvesting DarwinCore Records

One approach to estimating the total time to harvest all DiGIR data would be to develop an equation that assumed a linear relationship between the number of available DiGIR resources and the number of records in all DiGIR resources with the times available from the tests above. This relationship could take the form of:

Equation 3: $TST = NR * RST + NH * RH$

TST is the total request time, *NR* is the number of resources to search, *RST* is the time to search each resource assuming there are no hits on the resource, *NH* is the number of search matches, and *RH* is the time to request data for each match. There were only 8 matches for the *S. invicta* search and if we use its time of 3,721 seconds, we can obtain an estimate for *RST* of 5.841 seconds. If we subtract this value from the duration of all requests that contained matches we obtain a total time of 48,708 seconds to request the 57,164 records that matched our searches. This provides an estimate for *RH* of 0.8521 seconds.

Equation 4: $TST1 = NR * 5.841 + NH * 0.8521$

The current DiGIR network includes 637 resources, which gives a performance estimate of:

Equation 5: $TST2 = 2563 + NH * 0.8521$

It will take just over one hour to complete a search across all DiGIR providers if there are no matches. As the number of providers increases this number will increase to over 1.5 hours for 1000 providers and over 16 hours for 10,000 providers.

If we use the current number of available DiGIR records of 66,030,608, we obtain the number of seconds to harvest all DiGIR data to be 56,267,244 seconds, or over 651

days! Almost two years is unacceptable to allow harvesting of all the data currently available. While this is only an estimate and the actual number may be significantly lower, it shows the importance of the mean time to obtain each record which is currently 0.8521. If we multiply this value times the estimated 1.5 billion specimens in the world we can obtain a time to harvest all specimen data that may potentially be online in the future (not allowing for additional specimens being collected) of 1,278,150,000 seconds or over 40 years!

We can obtain a desired harvest time per record by assuming that the data must be able to be harvested in one month. Dividing the number of records in DiGIR by the number of seconds in a month with 28 days (the minimum number of days in a month), or 2,419,200, we obtain a time of 0.037 seconds per record. In this study, the Missouri Botanical Gardens resource came the closest to this number with a harvest time of 0.098 seconds per record for the *Tamarix* search. Because some providers come close to the desired request time while others took much longer, this may indicate a potential problem with the provider software or the complexity of the DiGIR protocol.

Varying the number of records requested showed an exponential increase in the amount of time required to harvest data as the number of records increased (Figure 2-4). This is most likely caused by a problem in querying data from a database over a non-persistent connection. The first time a query is executed, to return the first 100 records, the database will search through the data until it finds 100 records and will then return the data. The second time a query is executed, to return the next 100 records, the database must go through the same 100 records that were returned in the first query and then find

the next 100 records. This creates a geometric series, which gives the performance equation:

Equation 6: $T = C1 * 0.5 * NR \wedge 2 + C2 * 0.5 * NR$

C1 represents the time to fetch each record from the database and *C2* represents the overhead required to return each record to the program that requested it. The tests showed a small increase in the *C1* factor over the Database Query of 1.017 seconds, for the Simple DarwinCore Request, but a large increase of 9.828 for the DiGIR Provider. This is a serious problem and is most likely the cause of the extremely long durations to harvest the database on the *S. vulgaris*. The increase in the *C2* factor over the Database Query was 6.765 for the Simple DarwinCore Request and 14.40 for the DiGIR Provider. Since this is a linear multiplier against the number of records, it is less of a concern unless the effect of the exponential is reduced.

The performance results showed serious problems with certain providers. The DiGIR protocol showed that it is more complex than is required for the features provided. The provider software supplied to DiGIR providers is also more complicated than is required. The database design can also significantly influence the performance of searches.

A potential solution could be a single web script page that could be made available to providers. This would be much simpler than the existing solution and would make customization by providers attractive.

The existing solution will significantly reduce the number of servers and the amount of supported by a factor of about 10, bringing the mean time to request a record to approximately 0.085 seconds. Caching search results on the first request can remove

the exponential component of equation 6 which should bring the time per record below the target of 0.037 seconds. Increasing the number of records requested, for providers that allow it, can decrease this value further (Figure 2-4). Since harvesting systems, including GODM, will only need to harvest the data in entirety about once a year, monthly updates will reduce the performance required further making harvesting from larger datasets than DiGIR possible. Because updating the client software and the protocol can take significant amounts of time, all efforts should be made to insure that the protocol and the provider implementations are simple and execute as quickly as possible given the requirements.

The key concern of the DiGIR protocol and other web service protocols under development are their potential performance. Simple HTTP "GET" interfaces such as WMS provide a very fast and simple system for retrieving data. These systems can be scaled to a large number of web service providers with little cost, complexity, or performance degradation. DiGIR is among a class of simple XML protocols and currently provides adequate performance for several hundred providers. If the number of providers greatly increases, the performance of harvesting data from DiGIR providers will degrade accordingly. Of even more concern are proposals that call for much more complicated protocols based on very complex ontologies such as the Global Invasive Species Information Network. These protocols will provide much slower performance than HTTP GET interfaces or simple XML interfaces and have the potential to make DiGIR system so slow as to be of no use to systems such as GODM.

Another concern is providing support for a single web development language such as ASP or Java. This specificity will limit the breadth of support for any protocol

and since both of these technologies are proprietary, the academic community will have little influence over their direction (i.e., if the company decides to discontinue the product, software built using the product will not longer be viable). By creating the simplest possible protocols that meet the data sharing requirements, we obtain the highest reliability, performance, and the broadest base of support because it is easier to maintain toolkits for multiple languages.

2.4.1.3 Quality of Records

The results of the searches show that there is some valuable data within the DiGIR network. Since this data are from herbaria and museums around the world it includes coordinates for organisms outside the United States. Further, since non-native species, such as the members of the Genus *Tamarix*, may still be invading new regions of the United States, these records may contain important information on the range of environs that *Tamarix* can live in which may not be represented by its current range within the United States. These coordinates should be included in modeling. This type of information may be even more important to model potential invaders that have either not reached the United States or are just beginning to invade.

The quality of the data was also effected by misspellings and missing data. These problems can be filtered automatically but we can also expect a certain error rate associated with misidentification and incorrect locations which cannot always be filtered. A quality control system will be needed to notify providers of potential problems.

2.4.2 Environmental data

The JPL site showed excellent performance across a wide range of raster sizes, resolutions, and locations. The data transferred in the first case amounts to over 500 Megabytes of data transferred in just over 30 minutes in 1,000 separate transfers. This shows a mean transfer rate of 277 kilobytes per second. This level of performance is not only more than adequate for harvesting it also allows certain datasets to be served on the fly. These results also show that if the DiGIR could approach JPL speeds we would no longer have problems harvesting the existing DiGIR data and could harvest effectively from a much larger dataset.

2.4.3 Implications for documenting, mapping and modeling systems

The implications that can be drawn from the results apply to GODM and other similar systems that wish to address documenting, mapping, and modeling invasive species or other organisms. A key issue in determining feasibility of documenting species is the type of documentation desired. GODM allows users to browse or search to find information immediately in an interactive web environment. This requires high-speed access to species occurrence data. The results show that to incorporate occurrence data from web services, GODM will need to harvest data on a periodic basis and store it locally on GODM servers.

GODM has already demonstrated that web services can be used to create maps on the web browser by having the client software request raster images directly from a web service using a protocol such as WMS. GODM addresses issues of web service performance, coverage, and quality by requesting rasters from web services on the client initially and then replacing them when a final raster containing all layers is available from

the host server. To create a Portable Document Format (PDF) file for download the host server must have access to all data. Because of the issues identified with raster data from web services, any critical layers need to be harvested and kept locally on GODM servers. This process has already been completed for several global layers and for the US NED rasters, which were also processed to obtain slope and aspect raster layers for modeling.

Since executing a complex spatial model can be a slow process, it may appear that modeling could use data requested directly from web services. The problems with this include: (1) harvested data before modeling will speed up the modeling process and reduce the number of harvesting events to 1 if the data are reused for other models; (2) separating harvesting and modeling will reduce the complexity of the system and make harvested data available to other operations; and 3) harvesting is required for other operations, like mapping. The key point for all three processes is that harvesting data that will be used more than once improves the efficiency of the system. The tradeoff will be in providing storage for harvested data.

2.5 An Example

Below is an example that examines the modeling capability using data from DiGIR and from the existing GODM database for the invasive species *Tamarix*.

2.5.1 Methods

2.5.1.1 Data Acquisition

Data was obtained for *Tamarix* points from DiGIR providers and from the GODM database. All DiGIR data for the genus *Tamarix* that was harvested previously were examined including 222 occurrences. There were 52 occurrences remaining after the

removal of points outside the continental United States, duplicates, coordinates with less than two digits after the decimal, and points for species of *Tamarix* that are not considered invasive to the United States. A total of 14,107 points for *Tamarix* in the continental United Sates were obtained from the GODM database. All these points met the requirements outlined above.

Environmental data layers for the continental United States was obtained in the form of raster data. These layers included, precipitation, temperature, MODIS EVI Range, MODIS EVI Mean, distance to water.

2.5.1.2 Data Analyses

Values for the *Tamarix* points were added to the GODM and DiGIR datasets for each environmental variable. Histograms of the environmental layers for the United States, GODM *Tamarix* data, and the DiGIR *Tamarix* data were created. These histograms were created to contain 100 categories with the range of the categories set by the maximum range of the United States data.

The histograms were examined to determine which layers could best be used to spatially model the environmental niche associated with *Tamarix*. The mean, maximum, minimum, and standard deviations for the best candidates were computed.

2.5.1.3 Modeling

I attempted to model the data using the available environmental niche modeling tools, DesktopGARP (DesktopGARP 2006) and OpenModeller (OpenModeller 2006). The only tool that produced a final output surface was DesktopGARP and it was only able to complete this for one environmental variable with default settings. The only

remaining alternative to complete this analysis was to create a model using other tools. C++ was used to create a simple "box" model similar to BioClim. This model takes the minimum and maximum values for each environmental variable and creates a surface with 0's outside the ranges and 1's within the ranges. Since the box model only requires a minimum and maximum for each environmental parameter, the values computed previously were sufficient for parameters.

2.5.1.4 Map Generation

Maps were created using the simple box model with the GODM *Tamarix* data and then the DiGIR data, and finally with both datasets.

2.5.2 Results

Tamarix occurrence was most frequent in areas of low precipitation in both the DiGIR and GODM datasets (Figure 2-10, Table 2-8). The DiGIR dataset showed occurrences higher than the GODM dataset. On examination these points were found to be in middle to eastern Oklahoma and may represent data missing from the GODM dataset.

The GODM dataset showed a wider variance in the frequency of occurrence in temperature categories but both datasets showed a strong preference for temperatures above five degrees C (Figure 2-11, Table 2-9).

The occurrence of *Tamarix* in both datasets roughly followed the distribution of range available from the MODIS EVI product (Figure 2-12). *Tamarix* occurrence was associated with low mean EVI values in the MODIS EVI product (Figure 2-13).

Tamarix occurrences showed a correlation to short distances to water, but the GODM dataset showed another spike at about 50,000 meters to water (Figure 2-14).

Maps were generated using the precipitation and temperature ranges for the DiGIR dataset (Figure 1-17), the GODM dataset (Figure 1-18), and the two datasets combined (Figure 1-19).

2.5.3 Discussion

Tamarix showed a preference for dryer areas, but not the driest. It also showed a sharp drop in occurrence at a temperature of about 5 degrees C. The lack of presence points for the driest areas and for the coldest areas may indicate a sampling bias as these are areas that are more difficult to sample. The MODIS EVI data showed some general patterns, but these data cannot be used effectively without more sophisticated modeling tools. The distance to water showed that most occurrence points were close to water, which would be expected from a riparian species. The distance to water raster was created with a Shapefile that did not include small streams. This could explain the spike at 50,000 meters indicating *Tamarix* occurrences next to small streams in the middle of large dry areas.

The three maps show variable distributions for *Tamarix* as we change the range of temperature and precipitation values. The DiGIR map showed the effect of a narrower temperature range but a higher precipitation maximum. The GODM map showed a broader distribution from the DiGIR map but missed areas in the east and in the northwest indicated by the higher precipitation in the DiGIR map. The combined map displayed the widest distribution.

The biggest surprise from this experiment was the ability for the DiGIR data to produce a reasonable occurrence map with just 52 occurrences. The second surprise was that since the DiGIR data moved the maximum precipitation value up, this small amount of data had a significant impact on the predicted range of *Tamarix* when added to a much larger dataset. This indicated that the GODM dataset still may be missing data to fully describe where *Tamarix* will grow. I would recommend creating a stratified random sampling for areas that represent environmental niches that are missing *Tamarix* occurrences. These areas could be sampled for absence data, which would allow for much greater confidence in these models.

2.6 Future Directions

Results show that, with improvements, a cyberinfrastructure for invasive species management is feasible. This section presents a potential architecture for such an invasive species cyberinfrastructure. There are several integrated components required for the Invasive Species Cyberinfrastructure (Figure 2-18). The web services represent types of functions that are provided to the other components through the Internet through HTTP web services protocols like those determined feasible in the previous section.

2.6.1 Web Services

Inspection databases would maintain both the information on species that ports of call should monitor, and the observations made of invasive species, and the treatments applied for their removal. A valuable feature would be to maintain a list of the potentially dangerous invasive species, especially those that have not invaded, where they may be coming from, and what type of pathway they may arrive through. This would

allow inspectors to prioritize their inspections and focus on particular species. Inspection data can have significant economic and political implications and should be maintained by the customs and trade organizations.

Species occurrence databases include herbaria, museums, invasive species databases, threatened and endangered species databases, and agricultural databases. These databases can make their data available for harvesting by GODM through web service interfaces such as DiGIR.

The type of modeling proposed by GODM will require high-performance computers. GRID systems are being developed that can provide performance at the level of 1 million personal computers over the Internet. The Invasive Species Cyberinfrastructure can take advantage of these systems and contribute additional computing resources.

NASA provides a series of web services to provide remotely sensed data from satellites and airplanes. In addition to JPL NED data, other raster web services may be possible with organizations such as Earth Resource Observation and Science (EROS) Data Center. This information can be used to focus monitoring efforts on habitats suitable for a given invasive species (e.g., *Tamarix spp.*, see Morisette et al. 2006a).

2.6.2 Web Sites

Web sites are required to allow users to interact with the cyberinfrastructure. I have separated web sites from "tools" to draw a distinction between traditional web sites running on workstations and laptop computers and the more mobile interfaces available in PDAs and cell phones even though these two groups of devices use the same web technology.

Inspection web sites would allow users at ports of call to review information on the invasive species they should be watching for and to report observations of invasive species. These web sites would also allow users to receive regular updates to the software and data contained in their inspection field tools. The inspection web sites could be managed by the same inspection agencies managing inspection databases discussed earlier.

Resource management web sites provide resource managers for parks, refuges, harbors, counties, forests, and other areas with the information and tools to help manage their areas. GODM is an example of a resource management web site targeted at helping resource mangers to control invasive species through the use of field tools to capture information, maps of species distributions, statistical analysis of data, and models of current and predicted distributions, and the most effective control strategies.

Other resource management sites may focus on projection of threatened and endangered species, the spread of pathogens, or on commercial management of fisheries. Each type of web site may provide a different set of tools to their target users while accessing data from similar web services. An example may be that a resource manager examines the areas that have higher risk of invasion from fire by overlaying fire location and intensity data from EROS LANDFIRE on a map of the current distribution of invasive species in GODM. Another resource manager who is actively fighting fires may view information on how to keep certain invasive species from spreading into recently burned areas on a site dedicated to fire suppression. Through web services these two web sites would be exchanging data to provide the most critical information to a particular type of resource manager.

Web sites directed at citizen scientists are also needed. Cyberinfrastructures provide the capabilities to fundamentally change our ability to perform science and to make the results directly impact citizens (Ellisman 2005). Studies that engage citizen scientists are also more likely to collect data relevant to local conservation and management issues (Danielsen et al. 2005) and citizen scientists may have access to lands restricted to professional scientists and may find non-native species not previously detected (Lepczyk 2005). The invasive species cyberinfrastructure is one potential example of where citizens all over the world can not only benefit from cyberinfrastructure science but they may also participate.

Science web sites are focused specifically at users who are interested in having greater control over the statistical process and greater access to data than most users. GODM provides access to scientists by allowing the download of all non-sensitive data and advanced tools for analysis.

2.6.3 Field Tools

Field tools are hand-held or vehicle mounted systems that allow users to gather information directly in an electronic format, removing the need to enter the data manually after they return to their office. These tools also allow field workers to take information from the various services and web sites into the field for faster reference. With the development of wireless fidelity (WiFi) systems and satellite communication these tools will eventually be linked directly to the Internet and may remove the need to return to an office to upload data and to obtain updates.

Inspection tools will need to be provided to inspectors at ports of entry that allow the inspectors direct interact with the inspection websites during inspections. The

inspection tools would be available through a web site that provides installation and update functions. The data from the tools would be added to the inspection database using cell phone and WiFi technology. The ultimate inspection tool would be a hand-held device that allows inspectors to scan a bar code on a container and, based on knowledge from the cyberinfrastructure on where the container came from and what it contains, immediately display how critical it is to inspect the container and how to look for potential invasive species. If an unwanted organism was found the tool could, through wireless technology, recommend treatment, record the occurrence through web services, to the database, and notify the port of origin. Since invasions are cheaper to prevent than to control after invasion, this tool could be the single most important component of the invasive species cyberinfrastructure.

Field surveys provide information on the current distribution of invasive species and over time are critical to predict the rate of range expansion and the effectiveness of controls. For both inspections and field surveys, web sites can access information in the databases and provide appropriate email alerts to notify concerned individuals and organizations about new invasions. Currently field tools for invasive species surveys include paper maps, paper forms, Geographic Position System (GPS) units, and a number of Personal Digital Assistants (PDA). The paper and GPS tools require the surveyors to manually enter information into database or web site forms after the survey is completed. Software such as EcoNab (Graham 2006) and Weed Information Management System (WIMS 2006) allow the user to automatically move the data from the PDA into a computer.

In the future, we can expect the tools above to be increasingly replaced by PDAs that include more sophisticated software for surveys. Enhancements will include species identification guides and the ability to create custom surveys. As cell phone interfaces become more pervasive, the PDAs will be able to download maps and identification based on the current location of the surveyor. While surveys are being completed the results will automatically be uploaded to online databases.

2.6.4 GODM's Roll in the Cyberinfrastructure

Because GODM's primary focus is on modeling current and future distributions, and the effectiveness of control strategies, the most critical data to GODM is the taxonomic identification of a species at a location on the earth at a particular date. In addition, the type and nature of control efforts at locations must be available to modeling. Since the focus is on spatial modeling, GODM will need to be able to allow modeling across multiple spatial extents. This may require GODM to be able to manage large raster datasets. GODM will also have to allow raster datasets to be added "manually" by administrators, automatically by users over the Internet for their own rasters, and through harvesting techniques. Because of the long time durations to harvest raster data, this will have to be done as a background task before modelers can use the data.

GODM currently provides the ability for users to enter an occurrence for an organism, and associated data through a form-based interface, by clicking on a map, or by uploading datasets composed of columns of data (text files and Shapefiles). This allows small amounts of data to be added over time. Because GODM has limited support and resources it would not be possible to support a large user base of users uploading data. Increased attention is needed for customer support staff a greater investment in

hardware and software development may be needed. The only alternative is for GODM to work with other systems through a web service interface to obtain data from a much larger cyberinfrastructure. This research has shown that it is possible to harvest large amounts of occurrence data and raster data from web services.

2.6.5 Benefits

Up-front database design may target specific field data needs, while eliminating extraneous variables. For example, an investigator may save time in the field by collecting latitude and longitude information, then allowing geographic information system algorithms to model slope, aspect, distance to roads, and distance to water. The time savings in the field could be used to collect more sites. The database design might integrate remotely sensed information on vegetation type and structure, climate data, or satellite data surrogates such as MODIS data to facilitate the development of probability maps of species distributions to target future survey efforts (Morisette et al. 2006b).

The role for cyberinfrastructure development follows from complex database development. Recent advances in computer performance, storage capabilities, and network performance are allowing more complex ecological models to be used and for the data and results to be shared across the Internet (e.g., www.NIISS.org).

The immediate benefit of early cyberinfrastructures development in ecological studies includes the advantages of high-performance computers versus desktop computers, large storage devices for remote sensing data and environmental data, and a network to connect them together. Additional computers or clusters of computers off-site may assist in processor intensive computations (e.g., large extent multivariate spatial models) by communicating with each other through service protocols to provide a higher

level of capability than any one computer or even organization could provide independently. An effective cyberinfrastructure also includes extensive web services. Web sites are required to make databases usable by different types of users (Maurer et al. 2000).

Cyberinfrastructures are rapidly developing in the life sciences (Arzberger et al. 2004), geophysical studies (GEON 2006), and ecology (NEON). They generally focus on providing access to high-performance computers and large datasets to scientists. The geosciences cyberinfrastructure is additionally targeted to provide modeling capabilities to scientists and non-scientists for natural phenomenon such as weather patterns and earthquakes.

> "Society can't get full value for its investment in science unless anyone desiring existing data actually gets them" (Maurer et al. 2000)

Commercial cyberinfrastructures integrate point of sale, credit card processing, and shipping services on the Internet. These services must be fast, reliable, and secure to provide users with a positive experience (Foster 2005). Because invasive species may spread with commerce and trade, the same attributes are paramount for invasive species cyberinfrastructure. Unfortunately, current capabilities are woefully inadequate and are largely unavailable to resource managers. There are a large number of electronic databases on invasive species (Crall et al. 2006), but most of them are not available online and many are still maintained only on paper. Even the electronic databases that are online are inaccessible to other servers to exchange the exact locations or abundances of

invasive species. The coarse-scale raster maps that are displayed on many sites are of little utility to most resource managers.

The Invasive Species Cyberinfrastructure will be used by scientists for research and by resource mangers to manage invasive species on a daily basis. Because of this, such a capability will require high-performance computing, large datasets, and be very responsive, secure, and reliable.

2.7 Conclusion

The most important result of this research is that it is possible to harvest biological occurrence data from a distributed network of servers. It also shows that performance can be affected by the implementation, as in the case of the metadata and searching for records which both took far longer to request that expected.

Specific changes to DiGIR implementation should include: (1) removal of the "Number of Records" field from the metadata file and making the metadata file static: (2) providing the ability to search without a defined number of characters to allow harvesting all data in the most efficient method; (3) having a single script file for each supported language that uses configuration files by default but can be optimized as needed; and (4) insuring that the DiGIR protocol is as simple as can be given the requirements.

To make biological cyberinfrastructures successful we need to insure that: 1) protocols must be as simple as possible to reduce the protocol factor in the performance equation; and 2) provider software must be fast (<1 second for a no-hit search, <0.01 seconds per hit data transfer). To achieve this, the provider toolkits need to meet performance requirements on common systems and languages and need to be as simple as possible to allow it to be optimized for other systems and languages.

The challenges for the Invasive Species Cyberinfrastructure include getting scientists to release their data to the public (Maurer et al. 2000), creating effective protocols for transfer of detailed invasive species occurrence and management data, and the maintenance and support for a technology knowing it requires long-term, stable funding. In addition, the existing commercial Geographic Information System server products do not meet the scaling and performance requirements of the Invasive Species cyberinfrastructure.

We have already made initial steps in this development. The National Institute of Invasive Species Science (NIISS), a part of the United States Geological Survey (USGS) and many partner agencies and non-government organizations have developed a survey database, resource manager web site and associated field survey tools. Scientists can download data for analysis or use online analysis tools. This web site is available at www.NIISS.org. Current efforts include developing the protocols required to extend and complete the invasive species cyberinfrastructure. As with any cyberinfrastructure, maintaining and constantly improving the system requires frequent communications between users, researchers, engineers, and support staff; and routine hardware maintenance as would be required for any commercial web solution.

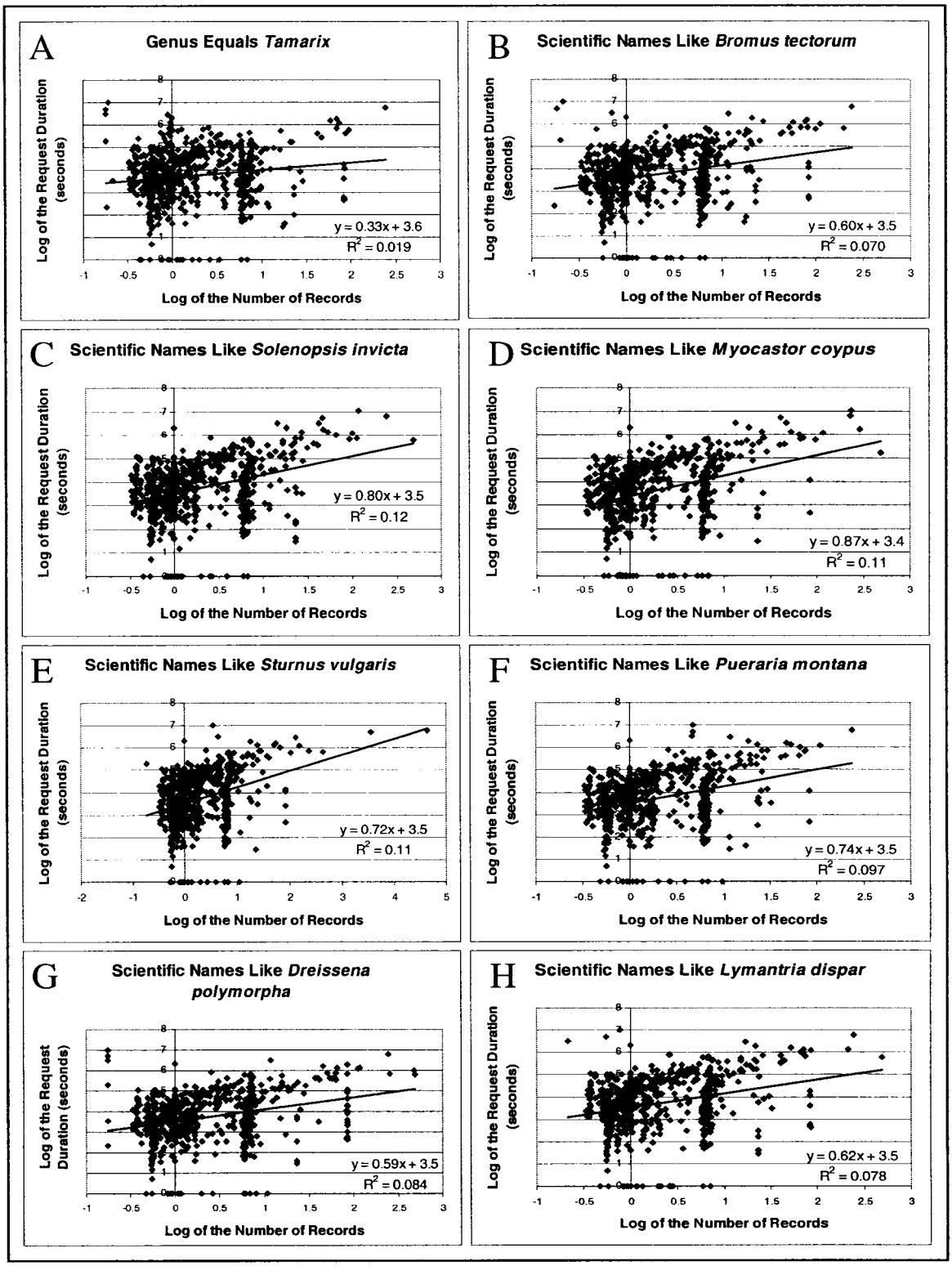

Figure 2-1. Request times in seconds versus the number of records in each resource. Each axis is a log scale while each chart is for a different harvest event for a given taxonomic group.

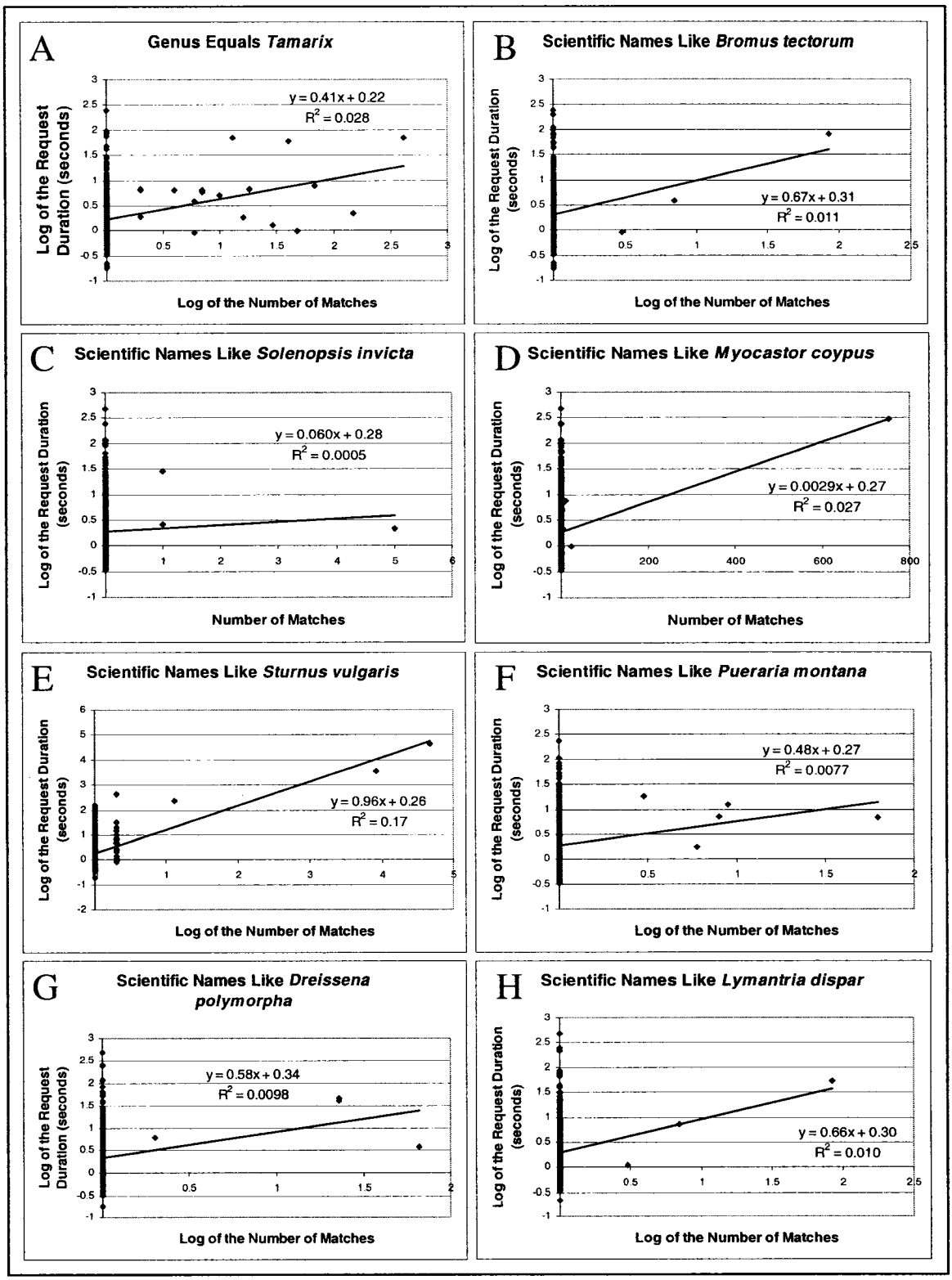

Figure 2-2. The time (in seconds) to search versus the number of records contained in each DiGIR database. Each axis is a log scale while each chart is for a different harvest event for a given taxonomic group.

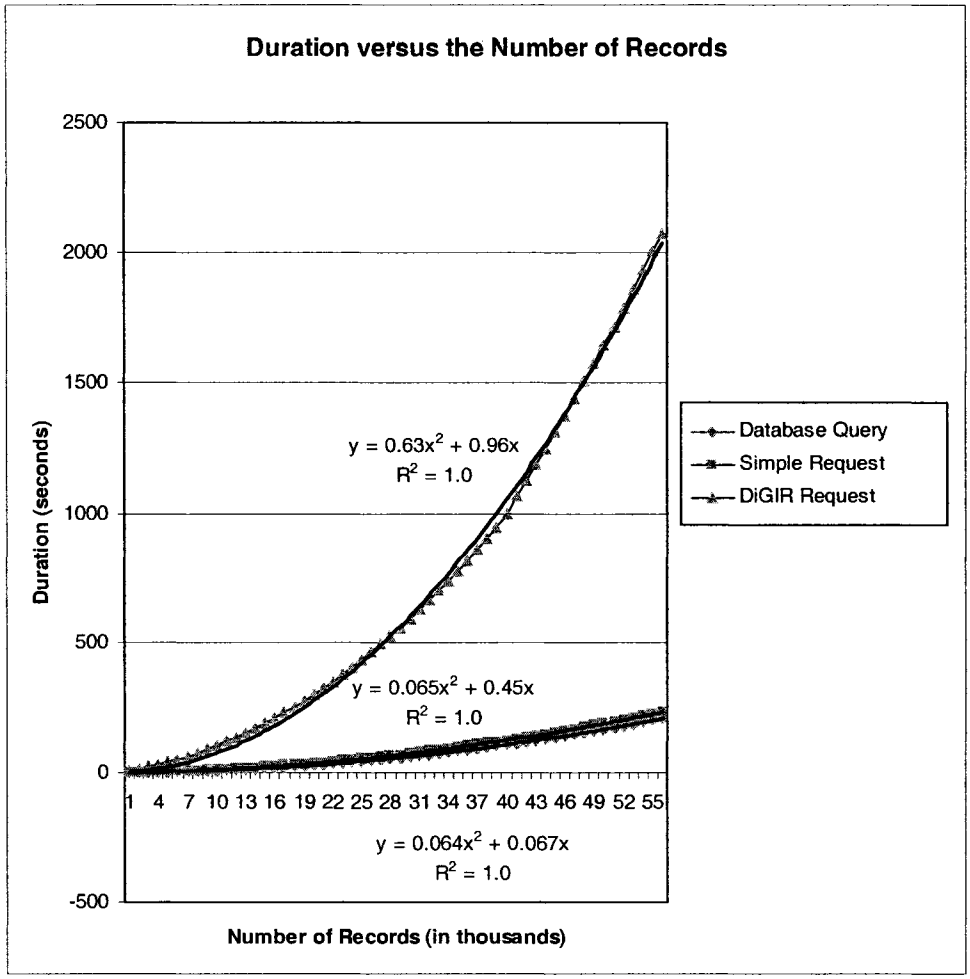

Figure 2-3. The time required to request data from DiGIR providers using the three different methods. The bottom line in diamonds is for a direct database query while the line just above it is for a simple DarwinCore request. The top line is using the DiGIR provider.

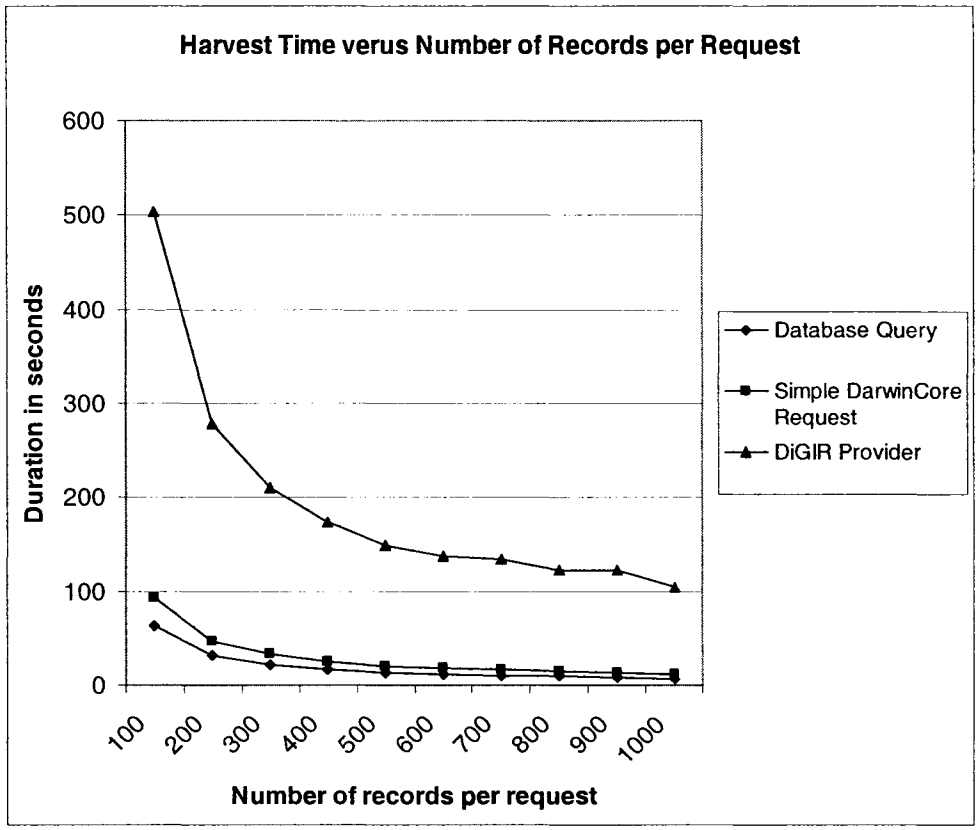

Figure 2-4. Harvest times with varying number of rows per request for the same query on the same system with a database query shown with diamonds, a simple DarwinCore request with squares and a DiGIR provider request with triangles.

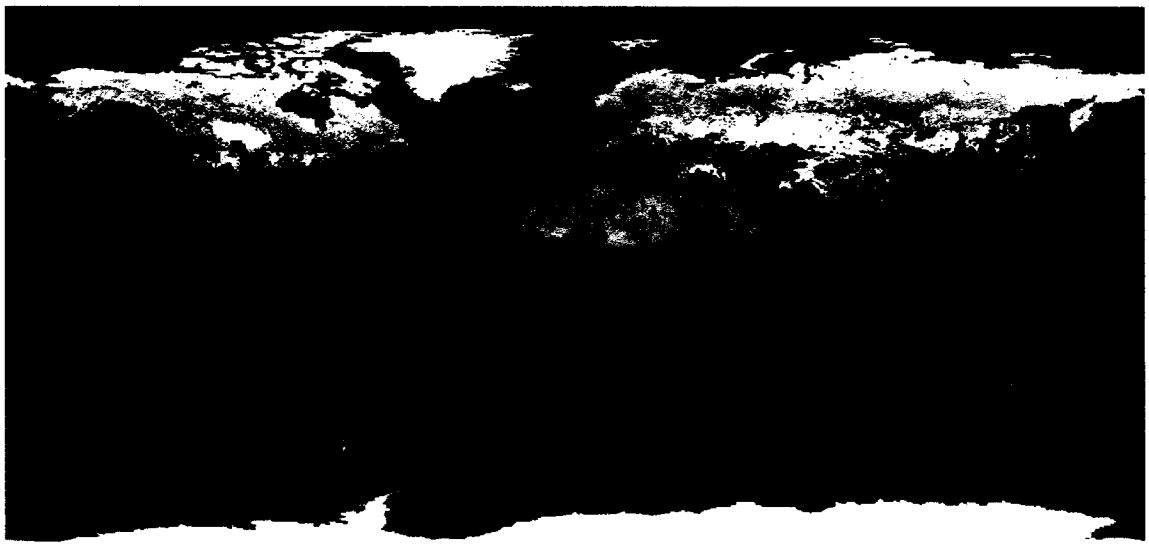

Figure 2-5. Map of the locations for organisms in the genus *Tamarix* from the DiGIR network.

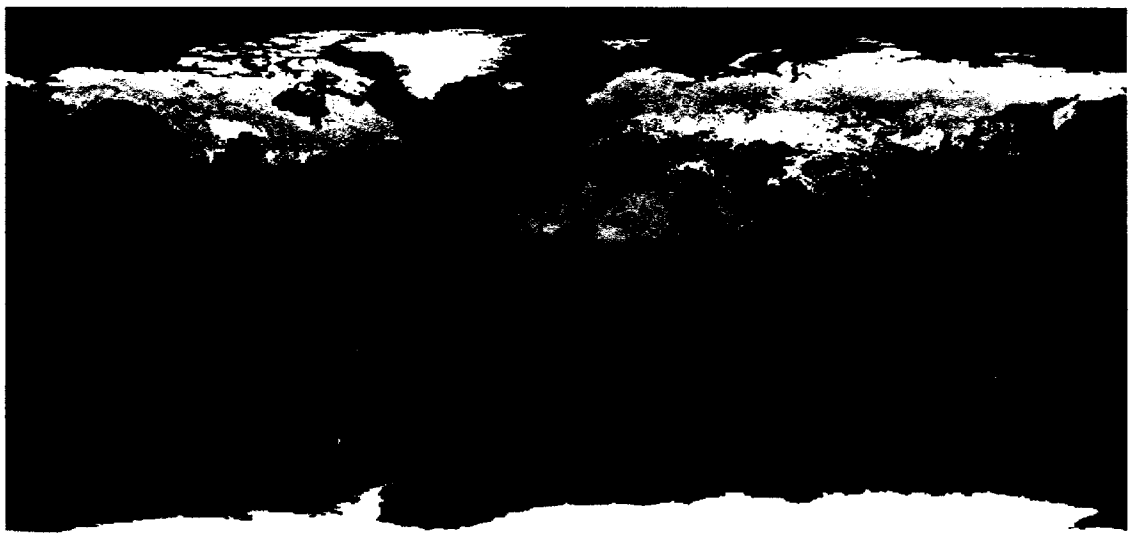

Figure 2-6. Map of the locations for organisms in the genus *Tamarix* that are currently in the GODM Database.

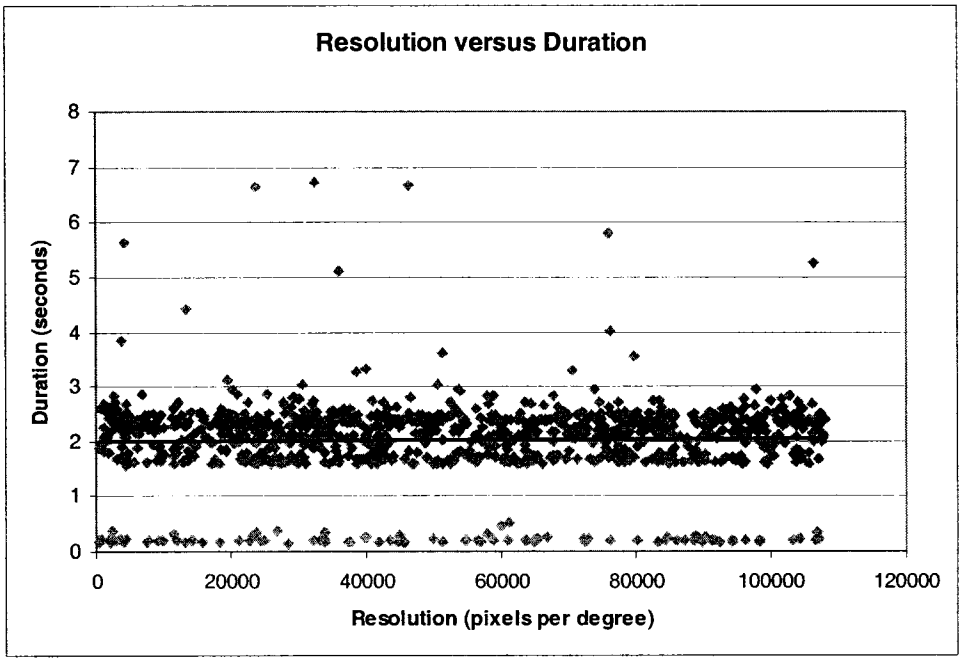

Figure 2-7. Time to acquire a 500x500 pixel, 16-bit signed, single band, raster from the JPL WMS web service plotted against the resolution in pixels per degree. Points near 0 duration correlate to failed transmissions.

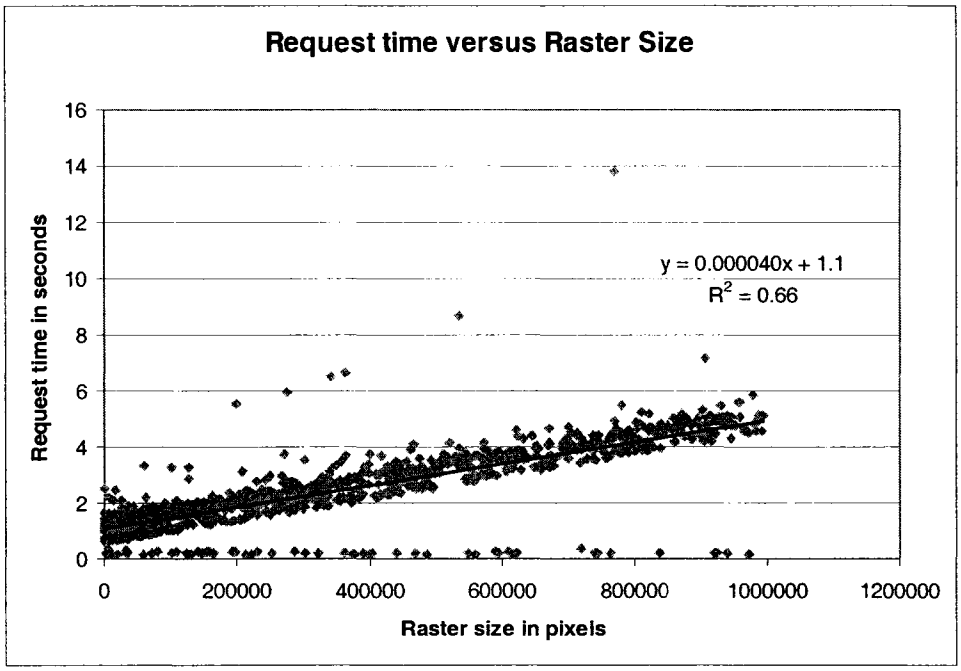

Figure 2-8. Request time for rasters of various sizes from the JPL web site including failures. Rasters were from the NED layer and were requested as GeoTIFF files.

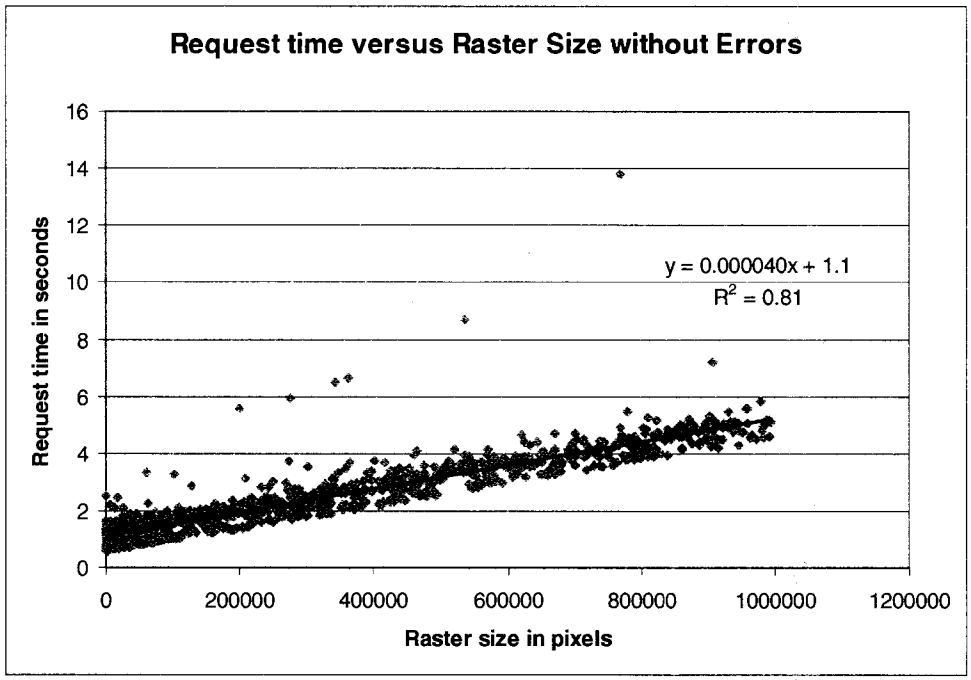

Figure 2-9. Request time for rasters of various sizes from the JPL with the failed transfers removed.

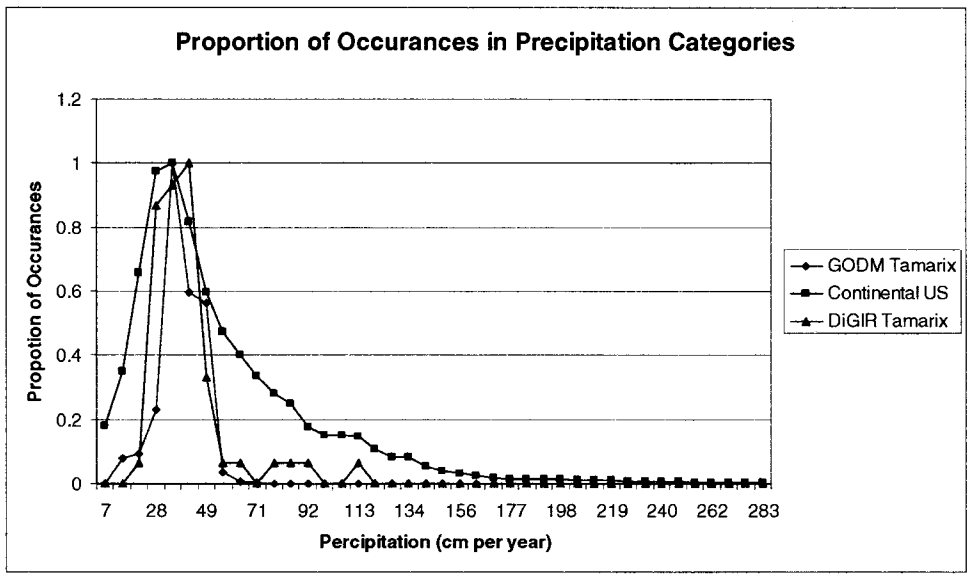

Figure 2-10. Histogram of the occurrence of *Tamarix* plants across the range of precipitation available in the United States. The x-scale was reduced from 707 cm per year to 283 cm per year show to more detail in the graph as there were only a small number of areas of rainfall above 283cm in the United States and no occurrences of *Tamarix* within them.

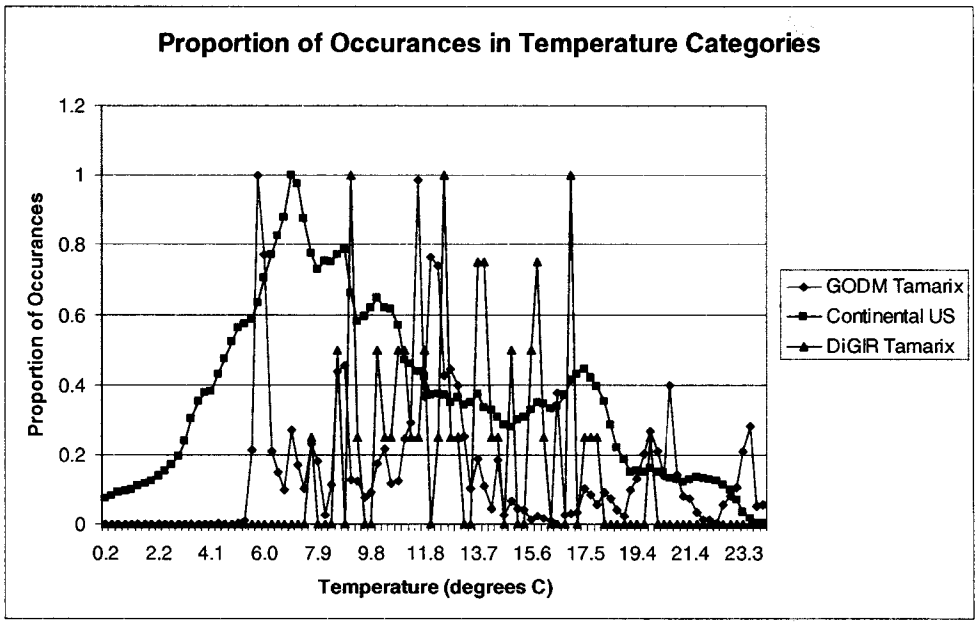

Figure 2-11. Histogram of the occurrence of *Tamarix* plants across the range of temperature available in the United States.

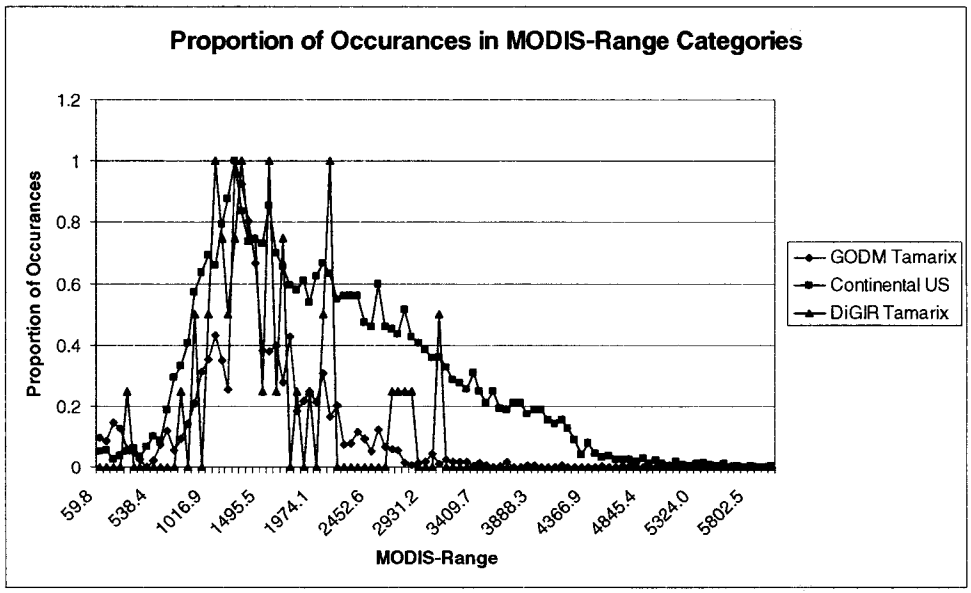

Figure 2-12. Histogram of the occurrence of *Tamarix* across values for the MODIS EVI product's range band.

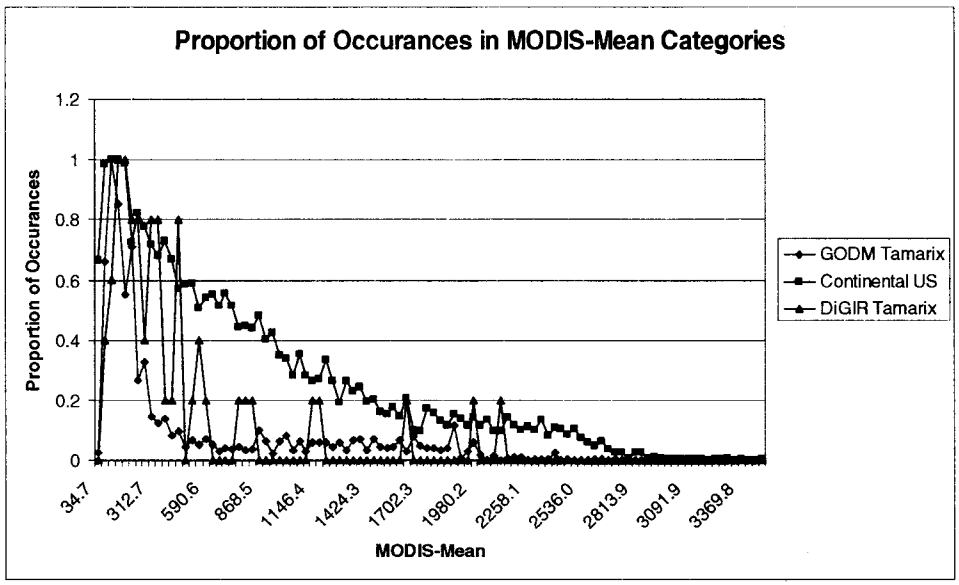

Figure 2-13. Histogram of the occurrence of *Tamarix* across values for the MODIS EVI product's mean band.

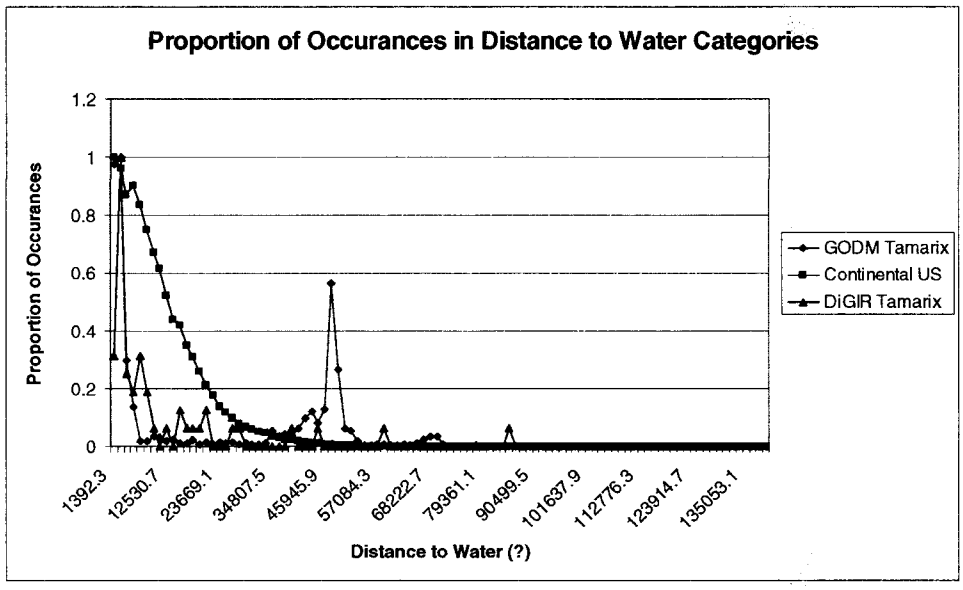

Figure 2-14. Histogram of the occurrence of *Tamarix* across categories for the distance to water.

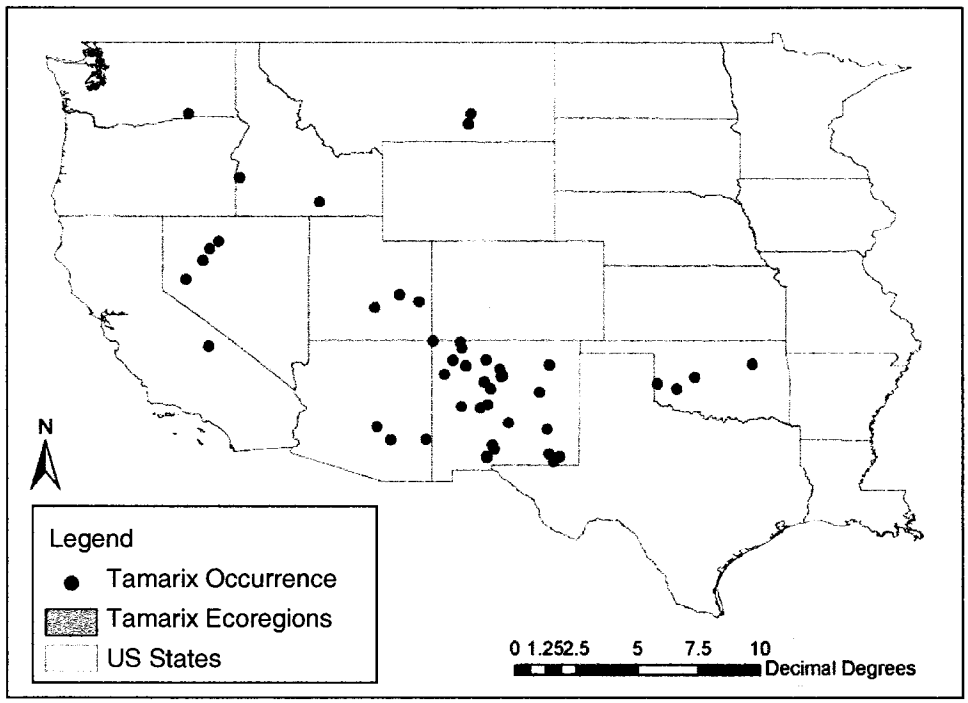

Figure 2-15. *Tamarix* ecoregions based on the precipitation and temperature ranges defined by occurrences from DiGIR providers.

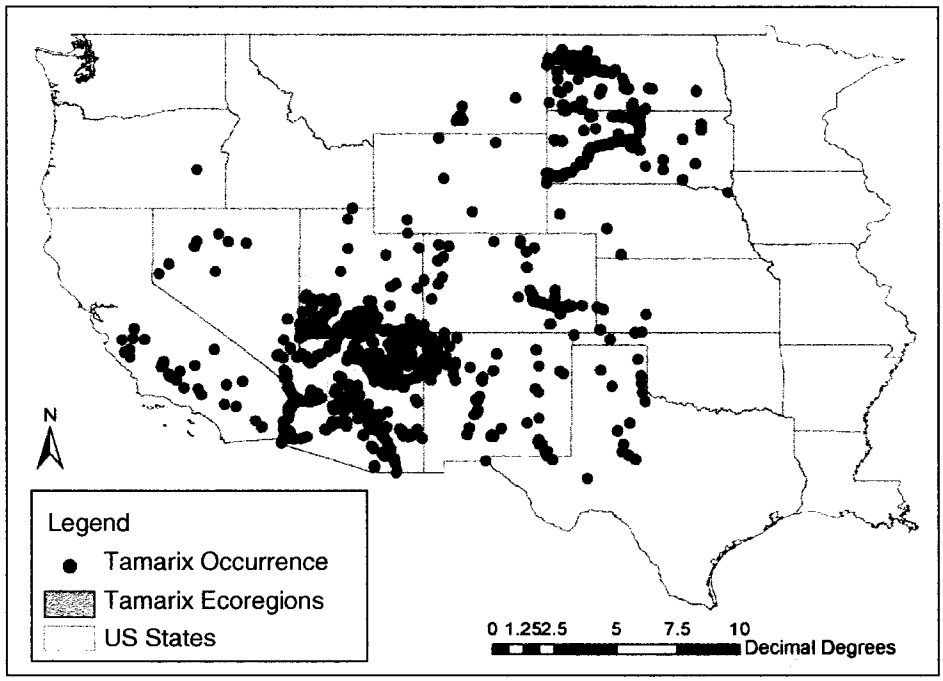

Figure 2-16. *Tamarix* ecoregions based on the precipitation and temperature ranges defined by occurrences from the GODM dataset.

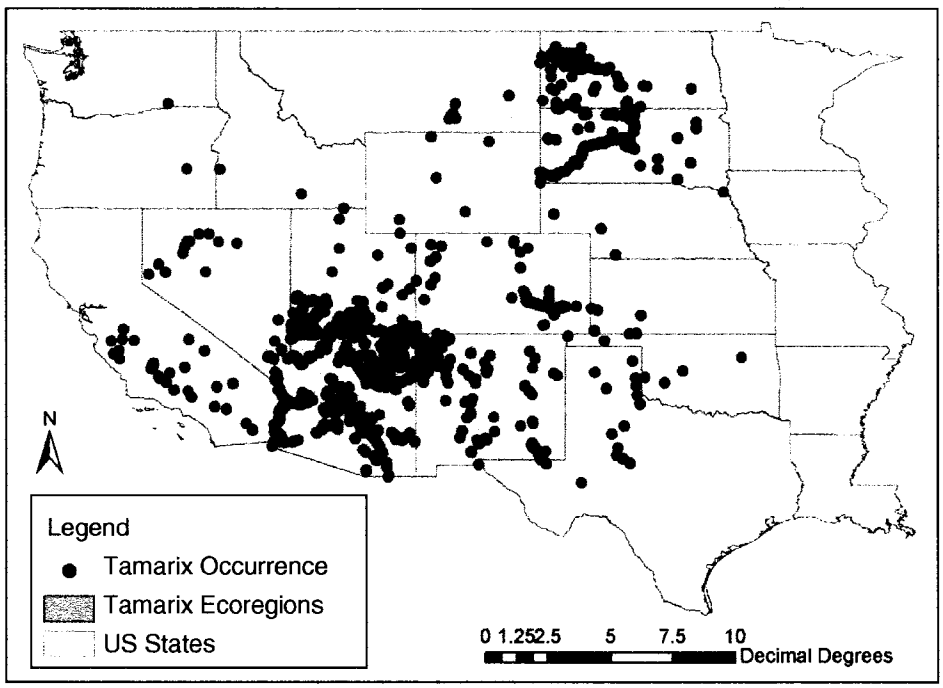

Figure 2-17. *Tamarix* ecoregions based on the precipitation and temperature ranges defined by occurrences from both the DiGIR providers and the GODM dataset.

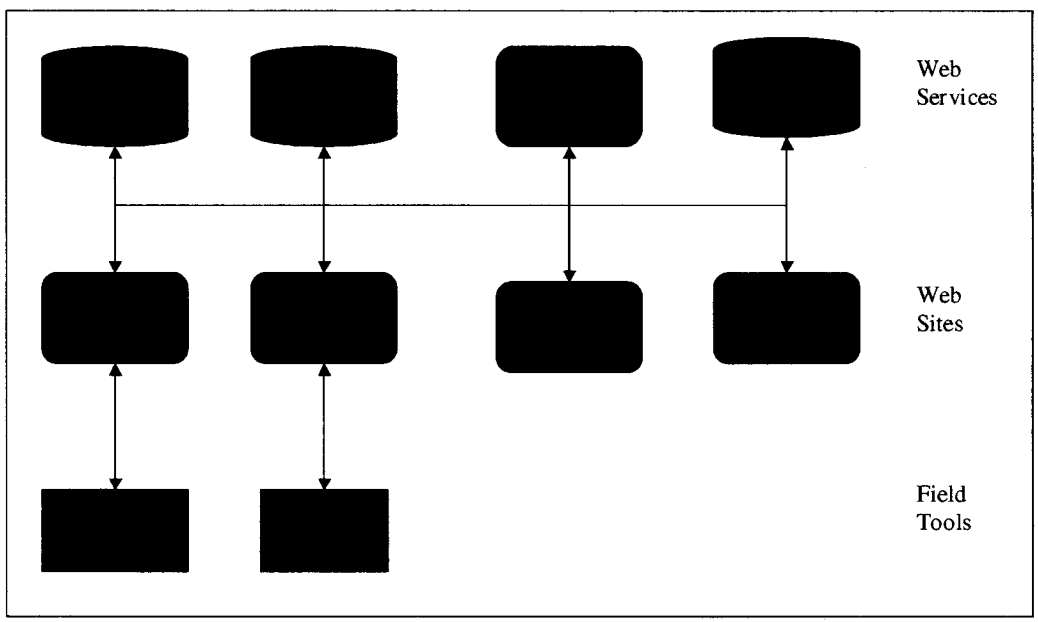

Figure 2-18. Architecture for a Cyberinfrastructure for invasive species

Table 2-1. Time (in seconds) to request information from 191 services on the DiGIR businesses from the GBIF UDDI database.

	Total Time	**Mean**	**Minimum**	***Maximum***	***Standard Deviation***
Request	56	0.48	0.44	3.4	0.28
Parsing	67	0.57	0.30	2.0	0.44
Total Time	123	1.0	0.75	3.7	0.51

Table 2-2. The first 3 rows contain the times (in seconds) to harvest all resources including the time to request the resources (first row), time to parse the response (second row), and the total time (third row). The last two rows contain the number of resources found in each service and the number of records found within each resource. Of the 663 resources, only 637 supported the DiGIR protocol.

	Total	**Average**	**Minimum**	**Maximum**	***Standard Deviation***
Request	2971	22	0.29	482	48
Parsing	363	13	0.0010	13.01	2.6
Total Time	3334	25	0.2967	482.3	48
Number of Resources	663	4.9	0	53	8.6
Number of Records	66,030,608	99,594	0	10,117,364	554,833

Table 2-3. Request time in seconds for various species for all 637 DiGIR based resources. The total number of records found and the mean, minimum, and maximum time to request a record across all resources.

Search Criteria	Total	Mean	Minimum	Maximum	Standard Deviation
Genus Equals *Tamarix sp.*	3,134	4.9	0.18	241	14
Species Like *B. tectorum*	3,900	6.1	0.17	241	17
Species Like *S. invicta*	3,718	5.8	0.33	117	23
Species Like *M. coypus*	4,066	6.4	0.34	481	27
Species Like *S. vulgaris*	51,098	80	0.19	43,788	1,7
Species Like *P. montana*	3,125	4.9	0.33	236	13
Species Like *D. polymorpha*	5,866	9.2	0.17	483	33
Species Like *L. dispar*	4,207	6.6	0.21	483	25

Table 2-4. Top 10 DiGIR resources by the total number of records in the resource with the search duration and number of matches for the criteria "ScientificName Like *Sturnus vulgaris*".

Service Name	Request Duration	Records in Database	Number of Matches
Avian Knowledge Network	3.52	10,117,364	0
The Swedish Museum of Natural History (NRM)	43,788	6,038,327	46,740
Avian Knowledge Network	3,594	5,201,311	8,300
Avian Knowledge Network	4.49	3,081,810	0
Instituto Nacional de Biodiversidad (Costa Rica)	134.93	3,031,345	0
National Research Foundation	0.00	2,111,770	0
Missouri Botanical Garden	18.20	1,966,000	0
European Environment Agency	54.66	1,646,269	0
Institute of Marine and Coastal Sciences, Rutgers University	50.20	1,374,234	0
Institute of Marine and Coastal Sciences, Rutgers University	25.10	1,281,125	0

Table 2-5. Top 10 DiGIR resources by number of matches for the criteria "ScientificName Like *Sturnus vulgaris*".

Service Name	Request Duration	Records in Database	Number of Matches
The Swedish Museum of Natural History (NRM)	43,788	6,038,327	46,740
Avian Knowledge Network	3594	5,201,311	8300
Australian National Herbarium (CANB)	231.6	706,766	13
Bernice Pauahi Bishop Museum	6.89	672,092	2
Museum of Vertebrate Zoology	434.7	660,312	2
Los Angeles County Museum of Natural History	6.67	538,209	2
Field Museum	20.30	440,821	2
Bird Studies Canada	4.50	408,413	2
Bird Studies Canada	2.65	271,569	2
University of Michigan Museum of Zoology (UMMZ)	4.79	203,570	2

Table 2-6. Top 10 DiGIR resources by search duration in seconds, for the criteria "ScientificName Like *Sturnus vulgaris*".

Service Name	Request Duration	Records in Database	Number of Hits
The Swedish Museum of Natural History (NRM)	43,788	6,038,327	46,740
Avian Knowledge Network	3,594	5,201,311	8,300
Museum of Vertebrate Zoology	434.7	660,312	2
Australian National Herbarium (CANB)	231.6	706,766	13
GBIF New Zealand	157.3	1,259,770	0
Instituto Nacional de Biodiversidad (Costa Rica)	134.9	3,031,345	0
National Herbarium of New South Wales	106.0	629,303	0
International Census of Marine Microbes	84.01	2,584	0
Siamazonia Provider	84.00	11,009	0
Herbario SANT, Universidade de Santiago de Compostela	84.00	13,319	0

Table 2-7. Multipliers for the coefficients for fitting a polynomial to the time required to request 55,068 records for the *S. vulgaris* with various methods. The local database query is set to 1.0 as a baseline.

	Exponential Multiplier	Linear Multiplier
Database Query	1.0	1.0
Simple DarwinCore Request	1.0	6.8
DiGIR Provider	9.8	14

Table 2-8. Numbers of records that matched the search criteria and the number that were rejected for various criteria. The last column shows the number of usable records for the GODM system.

Search Criteria	Resources With Matches	Matches	Number without coordinate	Number without TSN	Number without year	Number Usable
Genus Equals *Tamarix*	23	847	392	315	40	205
Species Like *Bromus Tectorum*	19	349	266	2	13	266
Species Like *Solenopsis invicta*	4	8	3	2	8	0
Species Like *Myocastor coypus*	15	810	807	718	730	3
Species Like *Sturnus vulgaris*	24	55,084	8,340	8	17	46,742
Species Like *Pueraria montana*	9	93	81	3	3	12
Species Like *Dreissena polymorpha*	9	119	44	2	5	74
Species Like *Lymantria dispar*	5	96	89	96	5	0
Total	108	57,406	10,022	1,146	821	47,302
Average	18	9,568	1,670	191	136.8	7,884
Percent of Total	NA	100	17.46	1.996	1.430	82.40

Table 2-9. Range of precipitation (cm) value for the various *Tamarix* datasets.

	Mean	**Minimum**	**Maximum**	**Std. Deviation**
DiGIR	34.3	13.4	105.7	17.4
GODM	31.7	7.7	69.2	8.7

Table 2-10. Range of temperature (C) value for the various *Tamarix* datasets.

	Mean	**Minimum**	**Maximum**	**Std. Deviation**
DiGIR	12.8	7.5	19.7	3.0
GODM	12.2	3.1	23.8	5.1

CHAPTER 3 AN ANALYTIC SOLUTION TO MAINTAINING ACCESS SPEEDS TO ARBITRARILY LARGE VECTOR SPATIAL DATASETS FROM INVASIVE SPECIES SURVEYS

3.0 Abstract

GIS web sites that allow users to upload new data can create a situation where there is a virtually unlimited amount of spatial data. As the data size increases, the performance of existing web mapping solutions would decrease proportionally. This decline is especially prevalent when the user is viewing areas of large spatial extents. We can derive a maximum rendering time by first understanding the nature of the data within GODM and then using analytic performance analysis for a solution that provides: (1) variable content based on scale; (2) places generalized vector data into a hierarchical grid structure; (3) uses an enterprise level relational database to manage the relationships between areas and their data within cells; and (4) renders "grid-pixels" at large extents instead of actual vector data. Performance can be further improved by rendering data directly from binary large data. Performance problems with finding spatial relationships can be removed by predetermining the relationships and storing them as relationships between areas in database tables. These developments will remove the constraints on the

amount of data a GIS-based web system can render to maps and improve access to data for other purposes such as analysis and web services.

3.1 Introduction

Multi-scale maps are already used in vehicle navigation systems (Zhilin and Ho 2004) and for vegetation mapping (Pun-Cheng et al. 2004). Recently multi-scale maps have become popular on the world-wide-web including MapQuest (MapQuest 2006) and GoogleEarth (GoogleEarth 2006). These web sites allow users to view maps of different areas by searching for locations and panning. Users can also zoom in and out to see maps at various resolutions. These maps include raster and vector spatial data. These web sites provide a set of data that is managed by the companies providing the service and significant efforts are made to process these datasets to insure the maps respond quickly for users.

3.1.1 Performance

One of the earliest techniques to improve the speed of map rendering was to limit the data handled to the area of interest that the user is viewing (Egenhofer 1994). Comparing the area of interest with the bounding box of all data in the dataset and eliminating all data that does not overlap with the area of interest can accomplish this. This operation requires a floating-point calculation on each item of data in the dataset and its speed will be based on the number of items in the dataset. This technique is provided by enterprise-level relational databases that support spatial operations such as Oracle Multi-dimension (Nieumwenhuijs 1995).

Another technique is generalization. Generalizing data to provide various levels of resolution, or levels-of-detail (LoD) has been used to create spatial databases.

Approaches included pre-computing different layers of generalization and generalizing data on the fly. Data can be generalized based on the data content or resolution (Cecconi and Galanda 2002, Prasher et al. 2003, Zhou et al. 2004). Work by Zhou and Bertolotto (Zhou and Bertolotto 2005) stands out from most other research on improving performance of vector data on the web by including performance evaluation.

Clustering, one type of generalization, is used to aggregate points in large spatial datasets. Approaches to clustering include partitioning, hierarchical methods, density-based methods, and grid-based methods (Pilevar and Sukumar 2005). Statistical Information Guide-based method (STING) is a project that using grid-based methods and indexing structures for points. STING places a grid of cells over points and then uses database indexing to quickly find the points that reside within a set of cells that overlaps with the area of interest (Wang et al. 1997). Section 3 of this paper extends this work to vector data within a hierarchical grid structure and adds the concept of "grid-pixels."

Progressive transmission of vector data between a server and a client on the Internet has been proposed (Yang 2004). The solutions in this paper are based on a "light-client" that will receive a raster from the server so progressive transmission is not appropriate.

Tiling data, also referred to as partitioning or gridding, has also been used to optimize access to spatial data (Robinson et al. 1995). The cells in the grid that overlap with the viewing area are identified, and then a table in a database is used to determine which spatial data items overlap those cells. This not only allows the floating point calculation for overlapping bounding boxes to be removed, but the cells can be organized as indexes in a relational database for immediate access. This method reduces the time to

access the spatial data from being based on the number of areas in the system to the amount of data within the area of overlap (Longley et al. 2001). This approach can improve performance for viewing small spatial extents but does not help when viewing large spatial extents.

There has been significant and related work in the field of data warehousing and online analytic processing (OLAP). This work is primarily based on non-spatial data, but in 2000 Stefanovic et al (Stefanovic et al. 2000) proposed an extension of this work for spatial information. The focus was on optimizing the access to commonly accessed objects rather than on developing techniques for limiting rendering time. Similar work by Pilevar focused on finding algorithms to detect new information from existing datasets or data mining (Pilevar, 2003).

While there is a significant literature on spatial databases, this research focused on extending the existing Structure Query Language (SQL) rather than using the characteristics of an existing relational database.

3.1.2 Data Commons

A new type of web site, referred to as data-commons, is emerging that allows for users to add data to online maps (Halpin et al. 2006). Examples include the Globe Program web site (Globe 2006) and the Global Organism Detection and Monitoring (GODM) system (www.niiss.org). If users were allowed to add data to these systems at even moderate rates, and then to view the data at large extents, the existing mapping solutions would begin to experience serious performance problems and eventually users would have to wait for hours for maps to render or would receive a "denial of service" message and would never be able to view a map.

Performance issues resulting from rendering complex vector data at a wide range of spatial scales and for different projections can be solved by examining the nature of the data to determine where problems may occur and then using relational database and GIS techniques to remove performance bottle-necks. This paper provides a solution to allow arbitrarily large vector datasets to be viewed quickly by users in a global web site. The content includes a novel combination of existing technologies and new approaches to solving the problem.

3.1.3 Global Organism Detection and Monitoring System

GODM, a joint project between the United States Geological Survey (USGS), Colorado State University (CSU), the National Aeronautics and Space Administration (NASA) and other organizations, is an online system for the rapid detection and response to invasive species threats around the globe.

GODM allows users to enter data on existing invasive distributions, view these distributions on real-time maps with other layers of interest, combine their data with other user's datasets, add remotely sensed data values and download the data for local analysis. Future additions to GODM will include the ability to model species ranges and to determine which types of control efforts are most effective in their area (Graham 2006).

Users can add their own spatial data from field surveys to GODM. These data may be in the form of points, polylines, or polygons. All these data are referred to as vector data. This data can be derived from hand digitizing on maps, Geographic Positioning Systems (GPS), and from processing remotely sensed data. The spatial data are accompanied by other information such as species identifications, individual organism attributes, types of treatments applied and environmental measurements. Users

also want to add political boundaries such as states/provinces, parks, and counties, and environmental boundaries such as hydrologic basins and to view these boundaries along with their data. Since users can add their own spatial data to GODM, and GODM is a global system, the amount of vector data contained in GODM is limited only by the storage space provided by GODM's computer hardware.

GODM is also required to aggregate and render data from global to local extents. Users must be able to view a map of all invasive plants in the entire world and then zoom in until they can view the invasive plants in their local area. Existing commercial GIS systems provide a linear access time as the quantity of data is increased. Since the amount of data will be continually increasing the system would quickly become unusable based on degrading performance, if existing software were used.

The features of GODM also require support for multiple spatial projections. Viewing data at global scales is typically done with geographic or spherical data. As the user zooms in it is common to convert the data to a different projection such as Universal Transverse Mercator (UTM). In analysis operations it is common to use data in an equal-area or equal-distance projection.

Most commercial web-based systems provide a fixed query capability to a non-relational database. Users of GODM will be allowed to create complex queries based on organism attributes, environmental attributes and the nature of treatments that have been applied to invasive species. This requires a solution where complex queries to a relational database must be executed very quickly and result in the desired set of spatial data. The way in which this data are structured and stored is critical to maintaining access speeds (Peuquet 1994).

Since GODM is a global system open to the public, the potential amount of Internet traffic can be quite high. The resulting performance problem could be addressed by continually adding computer hardware at significant expense. By insuring that the access speeds to vector data are as fast as possible we will be reducing this expense.

One feature of GODM is for the user to be able to obtain a list of the species in their country, state or other region. This requires the ability find all the survey data that overlaps with the selected area. This is typically a time-consuming task but must be done quickly in GODM.

GODM is expected to be an operational web site and already has a growing list of users but with a minimal budget for hardware and software development. The project needs to take advantage of existing or readily available computer hardware and software.

In summary, GODM must be able to quickly access and render complex vector data at a wide range of spatial scales and for different projections. The content of the data is derived from complex database queries on a dataset that is unlimited in size, constantly being updated by a global user community, and from a variety of sources. The amount of hardware available will be limited but the number of users will not be limited. Spatial relationships (Egenhofer 1994) will also have to be executed at very high speed on the same dataset. The data will also have to be made available for analysis and download.

3.2 Methods

To determine if it is feasible to maintain rendering performance with an arbitrarily large vector data set we will first define a topological data structure for the data, then

analyze the nature of the vector data that GODM has received from users, and finally, examine a performance analysis of the system based on various methods.

3.2.1 Topological Data Structure

The dominant format for vector data in the GIS industry is an ESRI Shapefile. Since Shapefiles do not contain a topological structure they duplicate portions of polygons that are common between them. Non-topological structures will also introduce "cracks" and "slivers" when being generalized. For these reasons a topological structure was developed.

3.2.2 The Nature of Vector Spatial Data

The nature of the complexity of the vector data that GODM will receive from users will be examined to determine the type of methods that will effective in maintaining rendering performance. This will include examining the number of areas of various types of geometries and the number of coordinates these areas contain. Then the size of the areas relative to the number of coordinates will be developed. Finally the density of coordinates for the areas will be determined on overlaying grids on samples of vector data from GODM.

3.2.3 Performance Analysis

An analytic model of the performance for rendering vector data will be developed. This performance model will be used as a representation of the performance we can expect from the actual system after taking into account the effects of computer hardware and programming languages selected for the system. The following methods

will then be used to modify the equation until a form of the equation is reached that shows the desired performance can be maintained.

1. Maintain spatial data in various projections
2. Provide variable content based on resolution
3. Create generalized layers at different resolution levels
4. Store generalized data in hierarchical grids
5. Optimize use of an enterprise-level relational database
6. Use a database-driven hierarchy for spatial relationships
7. Render spatial data directly from binary large objects (BLOBs)

3.3 Results

3.3.1 Topological Data Structure

The bulk of the data within the GODM system are spatial data from field surveys and other areas that can be described on the earth with two dimensional points, polylines, or polygons. A topological structure (Longley et al. 2001) is required within GODM to maintain the relationships between areas. To allow for the future representation of vertical and temporal (up to four dimensions) data and to allow for the storage of N-dimensional data from the results of spatial analysis, GODM vector data are maintained in an N-dimensional data structure. To maintain topology the structure was designed based on three-dimensional graphic constructs of vertices, edges, patches, and surfaces as shown in Figure 3-1. Vertices, edges, and patches are collectively referred to as the elements of the vector data describing an area.

3.3.2 The Nature of Vector Spatial Data

Currently the GODM system contains data from 61 projects including 37,945 field surveys, with the locations of 139,468 organisms representing 1562 different species (Table 3-1). The large number of non-survey polygons includes over 70,000 polygons representing the 24k, 100k, and 250k, quadrangle boundaries from the USGS topographic map series. We expect the number of survey areas to quickly bypass the number of other areas but the proportions of data should remain about the same. This indicates that the highest number of spatial data elements will be points but that the number of polylines and polygons will remain significant.

The actual number of coordinates in polylines and polygons can be over 10 times the number of coordinates contained within points (Table 3-2). Currently the majority of the data in GODM has been created from Geographic Positioning System (GPS) units. In the future we expect a significant amount of the data to be automatically created from remotely sensed data. These data can be more complex than data from GPS units. GODM will need to focus on problems of data management for polylines and polygons more than for points.

Histograms of the number of coordinates in each of the various types of polylines and polygons show that the survey polylines are dominated by polylines with a small number of coordinates (Figure 3-2). The same is true of the survey polygons. The non-survey polylines and polygons have a larger number of coordinates on average than survey data.

The areas covered of the various types of data are shown in Figure 3-3. The only projection that all the data in GODM is available in is Geographic (or un-projected).

Geographic data are not always appropriate for analysis but is appropriate here as Geographic is the projection for rendering data at broad spatial extents in GODM. There is a large range of sizes of the areas from 0, which are probably polylines with a single point, to over 100 degrees squared for non-survey areas including the political boundaries of states.

This information in Figure 3-3 could be used to determine a coordinate-density in the data based on the average size of the areas and the average number of coordinates as shown in Table 3-3 but these values are not an accurate representation of the coordinate density in the dataset because for large polygons the points are concentrated on the outside edges of the polygon. One example is the hydrologic unit datasets, referred to by hydrologic unit code (HUC). HUCs are derived from digital elevation data from remote sensors. This makes the data of very high complexity.

Overlaying a uniform grid of 100x100 cells over a hydrologic sub region showed that 97% of the cells for the hydrologic sub regions (HUC4) contained zero coordinates to 181 as the highest number of coordinates in a cell (Figure 3-4). The histogram of these cells showed that averaging the number of coordinates over the area of the polylines and polygons can be misleading as the coordinate density for the HUC4 computed from the total number of coordinates divided by the total area would be 690 coordinates per degree squared but the highest density within a cell in the grid is over 70,000 coordinates per degree squared. This difference indicated that the 3,102,000 million density may be low by two orders of magnitude.

As the user base of the GODM web site increases, we can expect the addition of over 10 million survey polylines and polygons each year. Data that are derived from

field GPS units will contain small polylines and polygons, while those derived from remote sensed data will be similar to the HUC data and will contain larger numbers of coordinates. This indicates GODM will be required to manage large numbers of very small, very complex polylines and polygons. The performance analysis needs to resolve the problem of rendering this large quantity of data when the user is viewing large spatial extents.

3.3.3 Performance Analysis

Performance can be analyzed by examining the different steps that are required for a computer to complete a task. For rendering vector datasets a performance equation can take the form:

Equation 1: $T1 = P * N + L * N + R * N$

Which can also be written as:

Equation 2: $T1 = (P + L + R) * N$

$T1$ is the total time to render a vector layer, P is the time to project a single coordinate, L is the average time to load a coordinate, R is the average time per coordinate to render a vector element, and N is the total number of coordinates in the dataset. This equation does not consider differences in rendering time based on length of line segments between points and filled versus unfilled polygons. Equation 2 shows that as more data are added the time to render increases linearly without limit.

One of the first techniques to apply to increase performance is to only operate on vector elements that overlap with the area being rendered. The problem is that users typically zoom in and out of a map based on a linear scale but, since this scale is applied to a rectangular area, the area changes as a square of the scale. Thus, when we

incorporate the zoom into the equation we add $Q * NA$ where Q is the average database query time per area in the database and NA is the number of areas defined and we replace N with $ND*Z$ where ND is the density of points within a given unit of area and, Z or the "zoom level", is the inverse of scale and is expressed as map units per pixel.

Equation 3: $T2 = Q * NA + (P + L + R) * ND * Z \wedge 2$

Equation 3 shows that as the density of data increases: (1) the time required to render the data increases linearly and as the user zooms out to lower resolutions and; (2) the time to render increases exponentially. The equation also shows that as the number of areas increases the time to query using linear a search increases linearly. To achieve the goals of this paper these three effects must be constrained to some domain.

3.3.4 Maintain Data in Required Projections

Converting between spatial projections, represented by factor P in Equation 3, can be time consuming. Projection on the fly can be performed using grids of projected points but this reduces spatial accuracy and is inappropriate for analysis. For this reason GODM will maintain data in both of the required projections Geographic and UTM.

Another projection problem that occurs is viewing data near UTM boundaries. The UTM projection system divides the world into 60 vertical strips where each strip is six degrees across. If the user is viewing data near one of the boundaries they may need to see data that is in another UTM zone. Polygonal and polyline data may also extend beyond a single UTM zone. For this reason, all data are maintained in all the zones that it overlaps with and one additional zone on either side.

Maintaining data in all required projections removes P from the equation and adds NP, the average number of projections required, to the query portion of the equation:

Equation 4: $T3 = Q * NA * NP + (L + R) * ND * Z \wedge 2$

Since we will be maintaining data in geographic and an average of 3 UTM zones, *NP* can be replaced by the value 4 giving:

Equation 5: $T4 = Q * NA * 4 + (L + R) * ND * Z \wedge 2$

3.3.5 Provide variable content based on resolution

Existing web-based geographic systems provide varying levels of detail based on the selected resolution to reduce the value of *ND*, MapQuest is one example. When the user is zoomed out to the maximum extent they see only a very simple version of the world. As they zoom in, additional content is added including state outlines and cities. As they zoom in further major roads begin to appear. At the maximum zoom level the user can see all the streets in their area of interest.

Within GODM the user can select certain types of layers to include with their maps. Layers such as roads can show interstate freeways at low resolution and then add highways and local roads as the user zooms in. The same can be done for cities by showing major cities and then adding smaller cities as the user zooms in. The same is also true for hydrologic units (HUCs) and river systems. While this approach provides a maximum number of elements for certain types of layers, we do not have criteria that can be applied to other types of layers including all survey data and, thus, this approach cannot be used to reduce *ND*.

3.3.6 Create generalized layers at different resolution levels

Simplifying polylines and polygons, also called generalization, is a common technique for reducing the amount of data in vector datasets. This section does not

consider generalization of points also referred to as clustering. The algorithm is a based on the Douglas-Peucker algorithm (Douglas and Peuker 1973) .

The original data have more complexity than can be shown at the rendering resolution (Figure 3-5-A). Generalizing to 0.1 degrees per pixel shows only a change in the complexity of the natural outlines (Figure 3-5-B). Generalizing to 0.2 degrees per pixel shows noticeable line straightening (Figure 3-5-C). This image should only be used at 50% of the resolution shown so the degradation would not be noticeable. This trend continues with the other images showing that the level of generalization cannot exceed the rate the user is zooming out from an image without introducing some image degradation.

The number of coordinates in the edges reduces at a rate that is faster than the decrease in screen resolution, from 4.3 to 2.6 (Table 3-4). However, the screen resolution is being halved; the amount of area exposed is a multiple of four requiring a decrease in the size of the data by a factor of four. This indicates that to maintain the quantity of data as the user zooms out we may have to have a small degradation in image quality. The original number of vertices is 95 in the original Shapefile, which contains unconnected polygons. After conversion to a topological structure where the edges are shared by neighboring polygons, the number of vertices stays constant at 195. This number will be reduced by subsequent techniques.

Based on the results above, if we are willing to absorb a small degradation in image quality, can replace the *ND*Z^2* factor in Equation 5 with *MZ*, the maximum point density for all zoom levels.

Equation 6: $T5 = Q * NA * 4 + (L + R) * MZ$

As the user zooms out the value of *MZ* will increase as the number of areas included increases. No matter how much we generalize each area the user will still be able to view the entire world with all the areas being rendered. This means that *MZ* is a function of the number of areas in the dataset and must be replaced.

3.3.7. Store generalized data in hierarchical grids

A gridded version of the outline of the boundary of Rocky Mountain National Park shows that grid cells may contain a portion of the edge of the boundary or may be completely within the interior of the boundary (Figure 3-6). Vector data only need to be maintained for edge cells. Interior cells that are not contained within an interior cell at a coarser grid level (see Figure 3-7) need only be maintained as relationships with between the area and the cell (Table 3-5). Cells that are completely outside the boundary or are interior cells at a coarser grid resolution do not need to be considered.

A generalized and gridded version of the boundary at one half the resolution of Figure 2-10 showed the quality of the rendering was maintained while reducing the number of cells required to represent the data (Figure 3-7). The relationships between the patches in the boundary and the cells in the grid have also been reduced (Table 3-6).

Examining at the problem from the server, it may at first seem impossible to provide a fixed maximum rendering time for a GIS system with an unlimited amount of vector data. Looking at it from the client's perspective, the user will rarely have more than a view of 500x500 pixels in a red-green-blue (RGB) format that takes three bytes per pixel, resulting in a 750,000-byte image. This size implies that in situations where there is a small amount of GIS data, it would be efficient to render the data into a map, but when there were large amounts of data, it may be more efficient to render each pixel

in the map by tracing back to the GIS data. This is similar to "ray tracing" in three-dimensional graphics (Glassner 1990).

The key concept is that if the data for a given area fit within a grid cell and the size of the cell is near or below the size of a pixel, simply fill in the "grid-pixel" with an appropriate color. As an example, if we have zoomed out from viewing the boundary shown in Figure 3-6 to the much smaller and simpler boundary in Figure 3-9 and then zoomed out again, we could render an 8 x 8 pixel version of the boundary by using the relationships in Table 3-5. These relationships are based on an 8x8 grid, and can be used for rendering, instead of the using spatial data. As we zoom out further we can use the relationships in Table 3-6, and then Table 3-7 and so on.

Rendering spatial data directly from grid-pixels means the number of coordinates within a polyline or polygon is no longer a factor in the performance equation and the equation is now based on the number of grid cells rendered for each area. The number of pixels is set by the resolution and size of the map rendering, and since this is a quantity we can control, we have a fixed maximum rendering time for any given area.

When the user is viewing data at high resolution, they will be seeing vector data in its true representation as points, polylines, and polygons. As they zoom out, smaller polylines and polygons can be rendered as grid pixels. When they are viewing the entire world, all of the survey data should have been reduced to grid-pixels, while some of the other vector data, like a country layer, will remain as vectors (Figure 3-10).

At very low resolutions we can set a maximum number of grid cells by replacing *MZ* with *NC*, the number of cells in the grid giving us a maximum rendering time of:

Equation 7 (low resolution): $T6 = Q * NA * 4 + (L + R) * NC$

NC has a maximum value of the number of pixels in the map being rendered. For high resolutions we replace *NZ* with *MH*, the maximum point density at high resolution.

Equation 8 (high resolution): $T7 = Q * NA * 4 + (L + R) * MH$

Since *MH* will be a much smaller value than the *NZ* term, equations 7 and 8 should provide an acceptable limit to the maximum rendering time with the exception of the *NA* term. This approach is also applicable for point-data as all the points in a single cell will become a single point when the appropriate low-resolution is reached.

3.3.8 Optimized use of an enterprise-level relational database

Equations 7 and 8 show that as the number of areas (points, polygons, or polyline based shapes) increases, the time to query using linear searches increases linearly. This is based on a floating-point query that examines the boundary of each area to see if it falls within the area of interest. Tree methods can significantly reduce the time, but to make the database access time have a fixed maximum we need to use direct indexing. Fortunately, with the data organized in a grid, we can use the rows and columns as indexes and replace the database query term, $Q * NA * 4$, with $Q1$ as the maximum time to access indexed data in the database.

Equation 9 (low resolution): $T7 = Q1 + (L + R) * NC$

Equation 10 (high resolution): $T8 = Q1 + (L + R) * MH$

This provides a maximum time for rendering, as equation 9 will typically dominate over equation 10.

A related issue is that GODM uses a relational database structure and users can create complicated queries. Performance is maintained by maintaining indexes for key tables within the database and the use of "views" for caching the results of complex

queries. This is well documented within the database literature on using Transaction Structured Query Language (T-SQL) (Celko 2005).

3.3.9 Use of database-driven hierarchy for spatial relationships

Determining spatial relationships is an operation that would scale linearly with the amount of data in the system. For this reason, when data are added to GODM, its relationships with certain "spatial hierarchies" are established between survey data and the other areas within the system (Figure 3-11). Nested areas can contain relationships with their enclosing area to allow larger areas to determine their relationship with surveys from their enclosed areas. Because these relationships are established as the users add data to the system, this step is removed from the rendering process.

Typically, generalized data must maintain the same relationships as the original data (Zhou et al. 2004). Pre-computing these relationships on the original data also removes this restriction.

3.3.10 Rendering spatial data directly from BLOBs

The internal memory structure for working with an N-dimensional topographic structure is relatively complex. Initializing this structure from BLOBs in the database was unexpectedly slow. To remove this performance problem the binary data format was modified to allow the data to be read for rendering and analysis directly from the BLOB without recreating the structure in memory. This reduced R in Equations 9 and 10 to more reasonable levels.

3.4. Discussion

Unique aspects of this research include; 1) database indexing to identify cells for all types of vector data, 2) the combination of hierarchical generalized layers with grids of cells, 3) the use of "grid-pixels", and 4) pre-computing and storing spatial relationships as hierarchical database relationships.

One of the most important aspects of this paper is the use of analytic performance analysis to guide the development of techniques to operate on large spatial datasets. Using analytic performance analysis can quickly eliminate approaches that may appear viable but that are far too complex to provide required performance.

The system designed above has been implemented in the GODM system with the exception of sections 4 and 7, which are planned on for mid-2007. The system has been created in a combination of PHP scripts for control functions, T-SQL for data access, and C++ for high-performance operations including all rendering. The system is running on standard IBM-PC compatible hardware with the Microsoft 2003 Server operating system, Microsoft Internet Information Server version 6.0, and the Microsoft SQL Server database. The PHP scripts are severed by the open source PHP library. Additional open source libraries include GDAL, OGR, JPEGLib, TIFFLib, FreeType, PNGLib and the GeoTIFFLib. The free version of the ECW file format library is also included from ERMapper, Inc.

With the vector data being projected into an average of four projections and then generalized at various levels, where the data size decreases by a factor of ¼ for each level, the overall size of the vector data can be estimated at six times the original data size. This may become a concern as the size of data increase.

3.5 Conclusion

This solution provides an innovative approach to using existing hardware and existing commercial software to produce a high-performance GIS system with a virtually unlimited ability to manage vector data. Each of the methods presented here solves a specific problem for vector spatial data. Together they provide a maximum access time to a specified set of vector data regardless of the amount of data; its resolution, or projection.

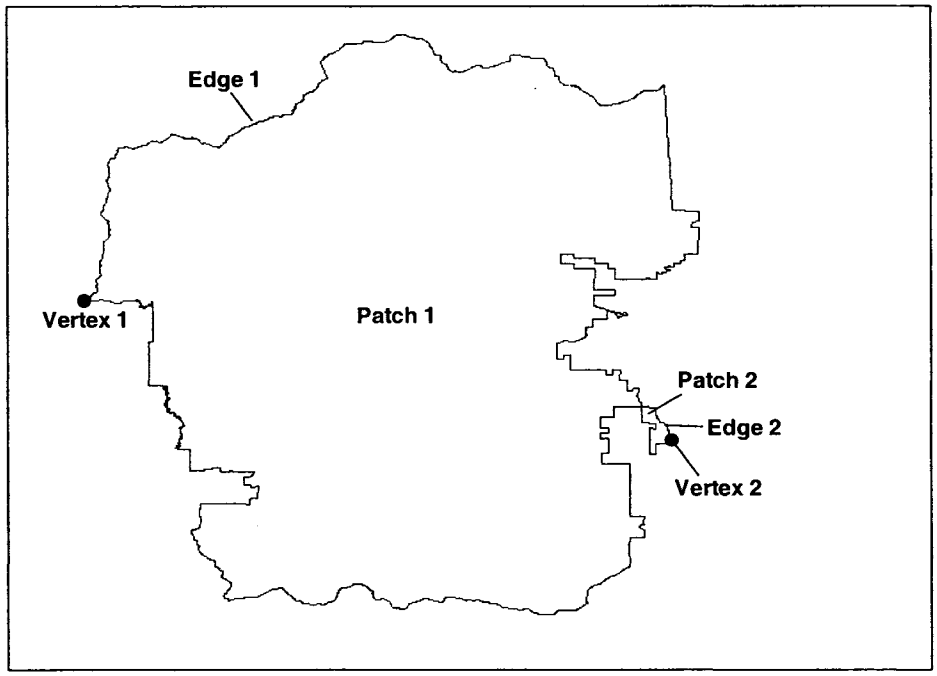

Figure 3-1. Vertices, edges and patches for a Rocky Mountain National Park, Colorado, United States. Data were taken from the Map of the World dataset provided with ESRI ArcMap.

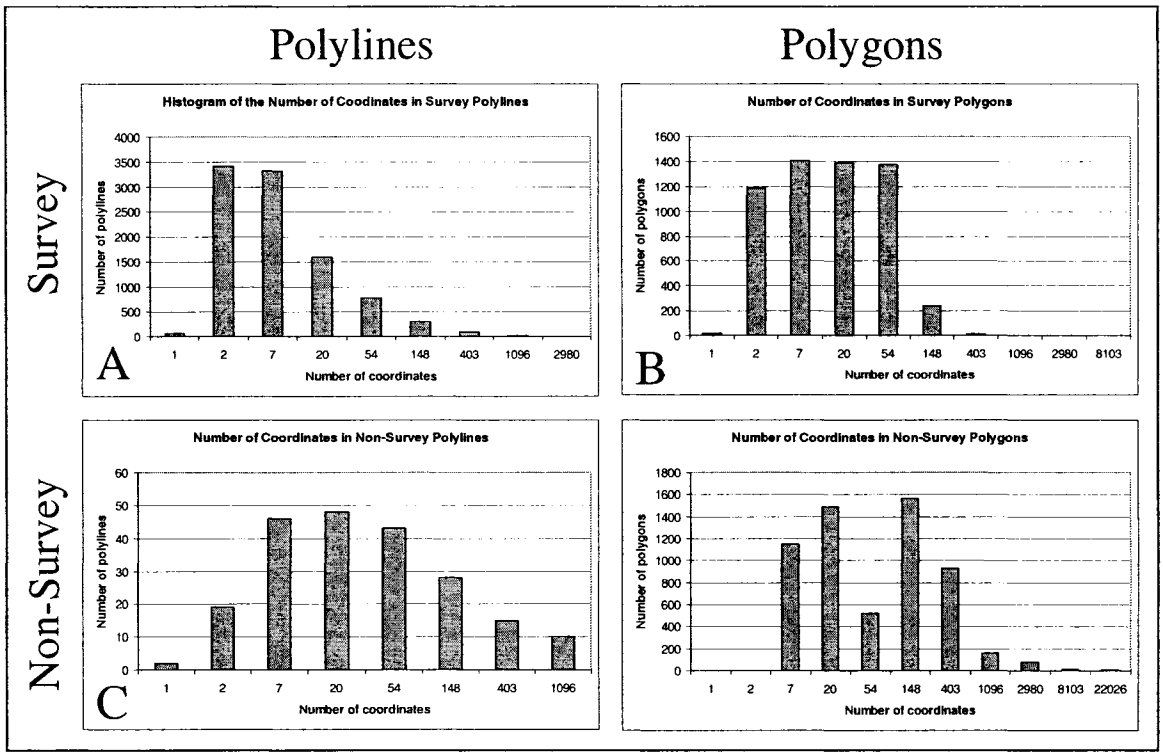

Figure 3-2. Histograms showing the number of coordinates in polygon and polyline data for surveyed areas and non-surveyed areas.

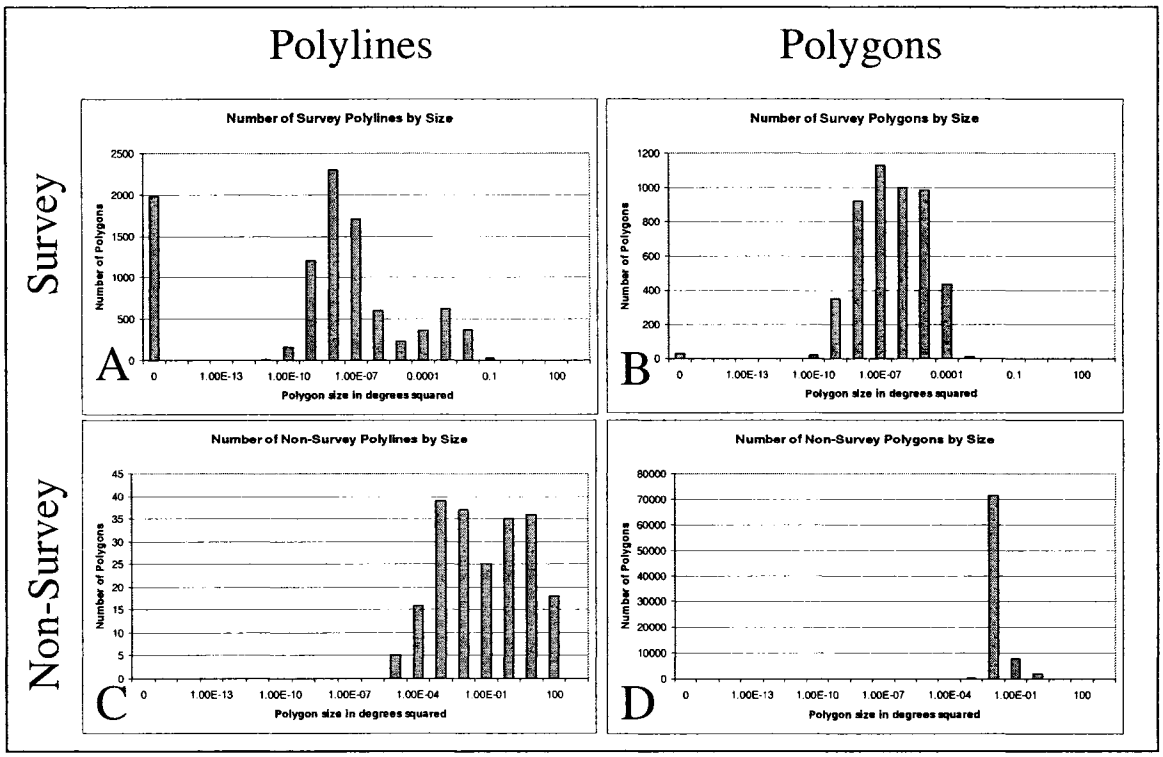

Figure 3-3. Histograms of the number of areas within quantized area sizes along a log-area continuum.

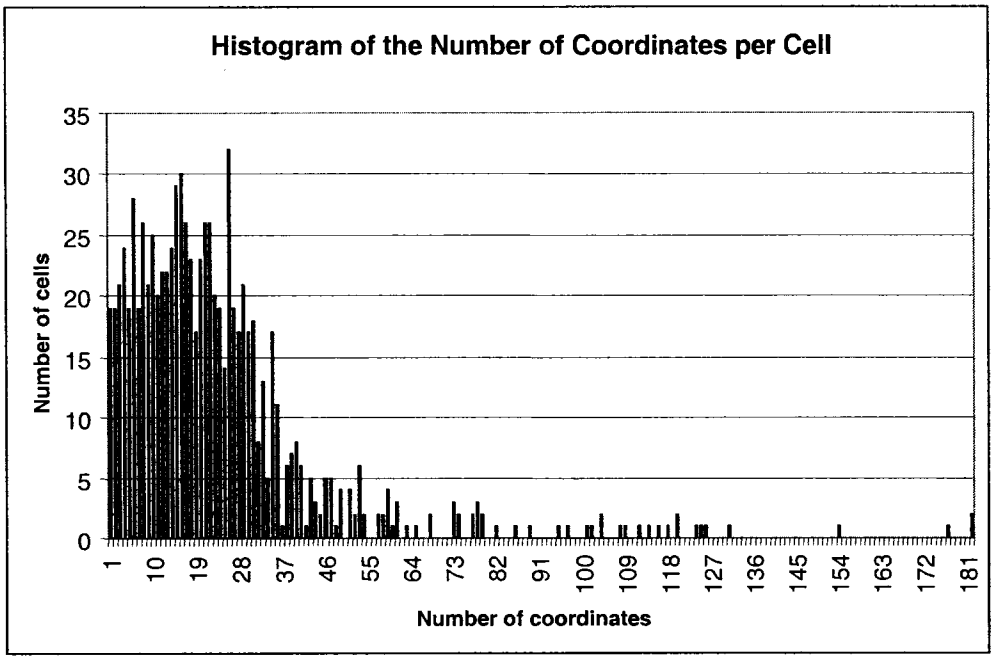

Figure 3-4. Histogram of the number of coordinates per cell for a 100x100 grid over a hydrologic sub region polygon with a cell size of 0.05 degrees. Cells with zero coordinates are not included.

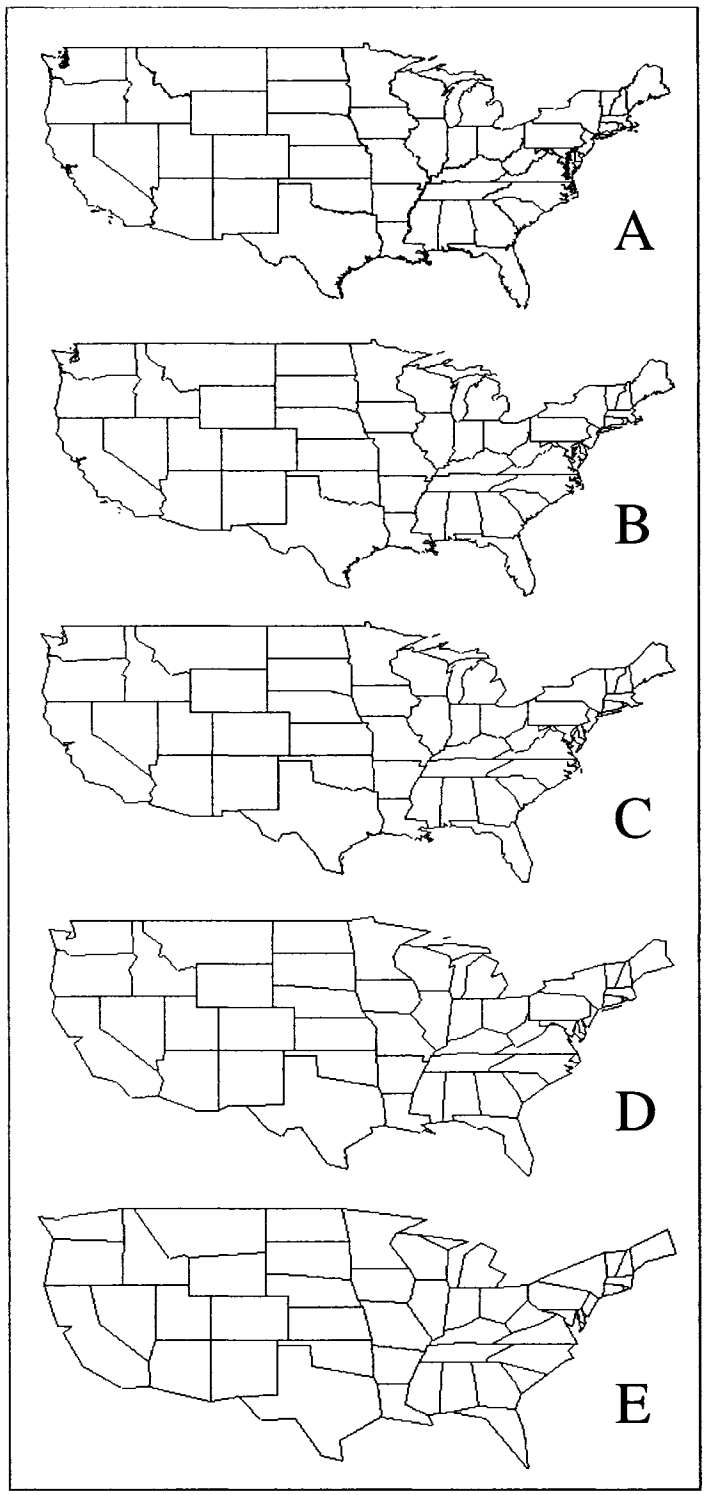

Figure 3-5. The outlines of the states of the United States at various levels of generalization rendered at 0.125 pixels per degree. A. Original image. B. Generalized to 0.1 degrees per pixel. C. Generalized to 0.2 degrees per pixel. D. Generalized to 0.4 degrees per pixel. E. Generalized to 0.8 degrees per pixel.

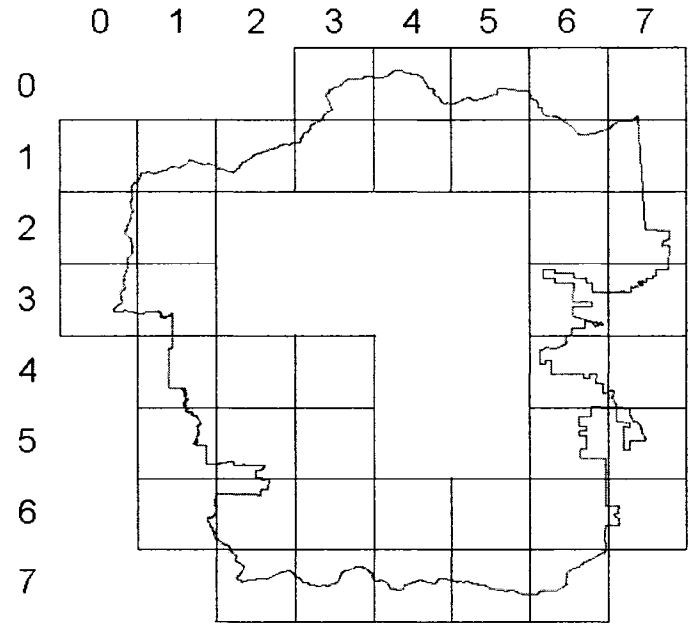

Figure 3-6. Grid cells that contain unique data when overlaid with the high-resolution boundary of Rocky Mountain National Park.

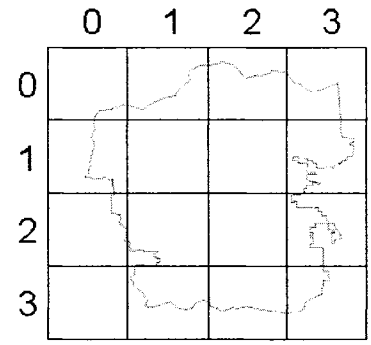

Figure 3-7. Grid cells that contain unique data when overlaid with a boundary of Rocky Mountain National Park that is one-half the resolution of the original

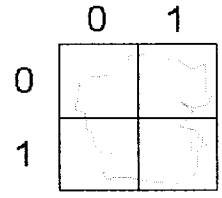

Figure 3-8. Grid cells that contain unique data when overlaid with a boundary of Rocky Mountain National Park that is one-quarter the resolution of the original

Figure 3-9. Low-resolution version derived from the relationships defined in Table 3-5.

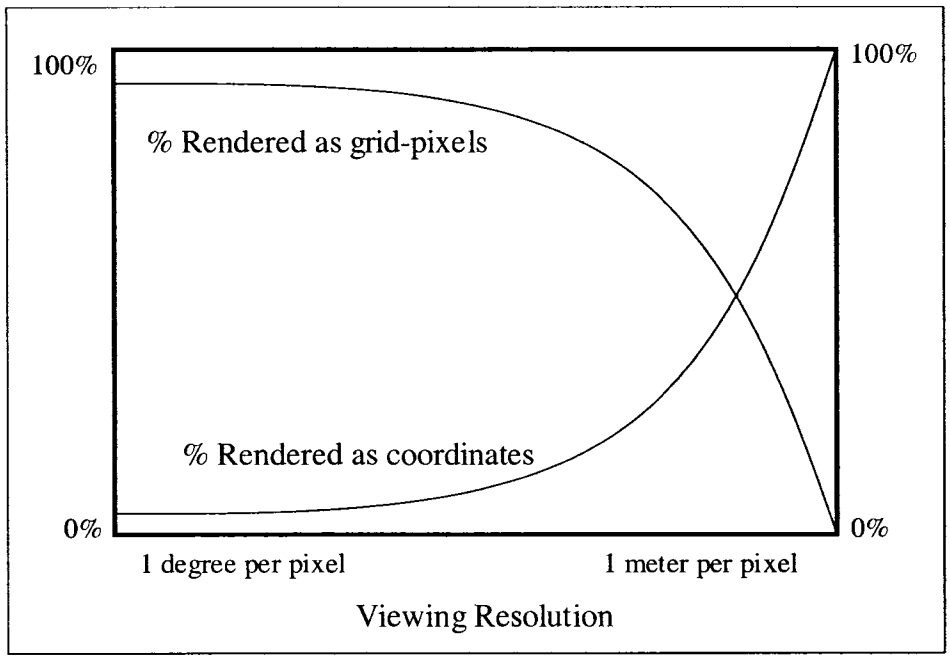

Figure 3-10. Relationships of the percentage of areas drawn using vector data coordinates versus grid-pixels. At low resolutions (1 degree per pixel), most areas will be drawn with grid pixels while at high resolutions (1 meter per pixel) all areas will be drawn with vector coordinate data.

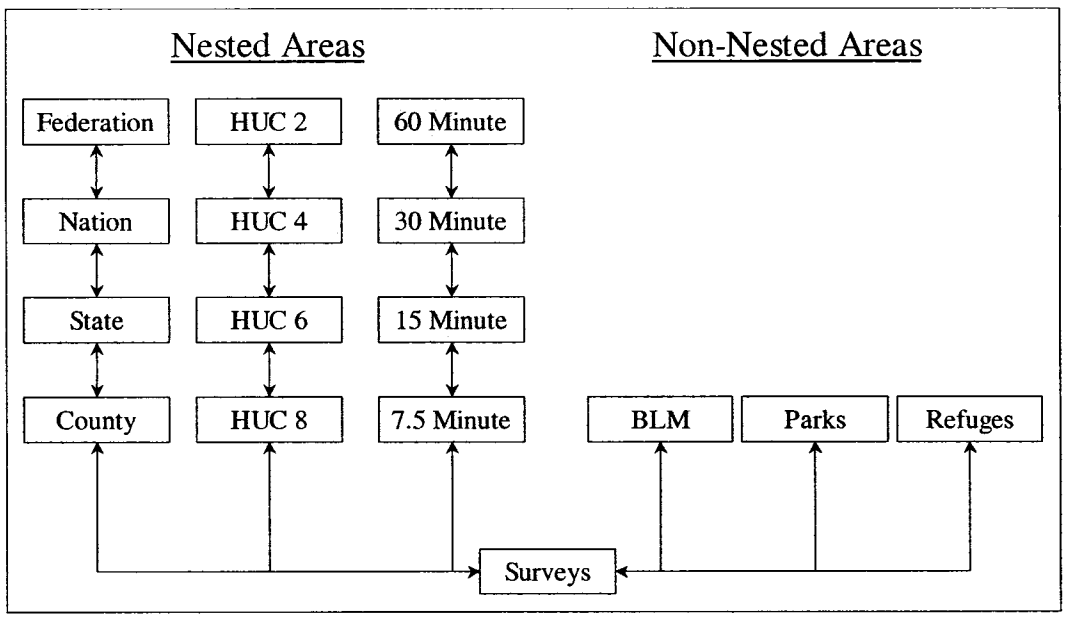

Figure 3-11. Spatial relationships between survey data and other areas. These relationships are established when the survey data or other areas are added to the system.

Table 3-1. The number of different types of geometry for spatial data based on survey and non-survey data.

	Survey	Non-Survey Areas	Survey and Non-Survey Areas
Points	24,057	48	24,105
Polylines	9,589	211	9,800
Polygons	5,623	82,039	87,662
Total	39,269	82,298	121,567

Table 3-2. Actual number of coordinates for each geometry type

	Survey	Non-Survey Areas	Survey and Non-Survey Areas
Points	24,057	48	24,105
Polylines	358,835	40,510	399,345
Polygons	268,629	82,039	350,668
Totals	651521	122597	774,118

Table 3-3. Average coordinate density for different types of data computed from the average number of coordinates divided by the average bounding box area.

	Average Bounding Area	**Average Number of Coordinates**	**Average Coordinate Density**
Survey Polylines	0.000768	37.42	47,970
Non-Survey Polylines	10.417	192.0	18.43
Survey Polygons	0.0000154	47.77	3,102,000
Non-Survey Polygons	0.06968	30.37	765.4

Table 3-4. Total number of points to describe the states in the continental United States. Each row represents a unique set of data with the last 5 rows matching the figures in Figure 5. Data are from the ESRI example Shapefiles from ESRI Data & Maps (ESRI 2001) .

Level of Generalization (degrees per pixel)	**Number of Edge Coordinates (percent of original)**	**Number of Vertices**	**Total Points (percent of original)**	**Degradation at 0.125 degrees per pixel**
Original	11,291	95	11,386	None
Edges Combined	6,729 (60%)	195	6924 (61%)	None
0.05 (A)	1542 (14%)	195	1737 (15%)	None
0.1 (B)	707 (6.3%)	195	902 (8%)	Complex areas become "less black"
0.2 (C)	297 (2.6%)	195	492 (4.3%)	Visible straightening
0.4 (D)	115 (1.0%)	195	310 (2.7%)	Oversimplified
0.8 (E)	36 (0.3%)	195	131 (1.1%)	Downright ugly

Table 3-5. Relationships between the patches within the Rocky Mountain National Park high-resolution boundary and the grid cells in an 8 by 8 grid including the element, column, row, and whether the edge is interior or contains edge data.

Element	Column	Row	Type
Patch 1	3	0	Edge
Patch 1	4	0	Edge
Patch 1	5	0	Edge
Patch 1	6	0	Edge
Patch 1	7	0	Edge
Patch 1	0	1	Edge
Patch 1	1	1	Edge
Patch 1	2	1	Edge
Patch 1	3	1	Edge
Patch 1	4	1	Interior
Patch 1	5	1	Interior
Patch 1	6	1	Edge
Patch 1	7	1	Edge
Patch 1	0	2	Edge
Patch 1	1	2	Edge
Patch 1	6	2	Interior
Patch 1	7	2	Edge
Patch 1	0	3	Edge
Patch 1	1	3	Edge
Patch 1	6	3	Edge
Patch 1	7	3	Edge
Patch 1	1	4	Edge
Patch 1	2	4	Interior
Patch 1	3	4	Interior
Patch 1	6	4	Edge
Patch 1	7	4	Edge
Patch 1	1	5	Edge
Patch 1	2	5	Interior
Patch 1	3	5	Interior
Patch 1	6	5	Edge
Patch 1	1	6	Edge
Patch 1	2	6	Edge
Patch 1	3	6	Interior
Patch 1	4	6	Interior
Patch 1	5	6	Interior
Patch 1	6	6	Edge
Patch 1	7	6	Edge
Patch 1	2	7	Edge
Patch 1	3	7	Edge
Patch 1	4	7	Edge
Patch 1	5	7	Edge
Patch 1	6	7	Edge
Patch 2	7	4	Edge
Patch 2	7	5	Edge

Table 3-6. Relationships between the patches within the Rocky Mountain National Park boundary at one-half the original resolution and the cells in a 4 by 4 grid.

Element	Column	Row	Type
Patch 1	0	0	Edge
Patch 1	1	0	Edge
Patch 1	2	0	Edge
Patch 1	3	0	Edge
Patch 1	0	1	Edge
Patch 1	1	1	Interior
Patch 1	2	1	Interior
Patch 1	3	1	Edge
Patch 1	0	2	Edge
Patch 1	1	2	Edge
Patch 1	2	2	Interior
Patch 1	3	2	Edge
Patch 1	1	3	Edge
Patch 1	2	3	Edge
Patch 1	3	3	Edge

Table 3-7. Relationships between the patches within the Rocky Mountain National Park boundary at one-quarter the original resolution and the cells in a 2 by 2 grid.

Element	Column	Row	Type
Patch 1	0	0	Edge
Patch 1	1	0	Edge
Patch 1	0	1	Edge
Patch 1	1	1	Edge
Patch 2	1	1	Contains

CONCLUSION

This research shows that it is possible to create a Global Organism Detection and Monitoring system that will allow the collection, integration and dissemination of information on the distribution of invasive species at a global level with immediate access for anyone with an Internet connection. The long-term impact of this type of technology can only be imagined.

Initially, government agencies and especially local personnel with existing relationships with Colorado State University will use GODM. This provides a focused group to help insure the right feature set is selected for long-term success. As GODM grows in acceptance, public outreach programs will raise awareness and empower everyone to join in mapping, monitoring, and controlling invasive species and restoring ecosystems. Users will include county weed crews, environmental organizations, ranchers, national resource managers, gardeners, and amateur naturalists. GODM will provide them with the tools to optimize their efforts to make the use of scarce resources by helping to set targeted priorities for invasive species, and highly vulnerable habitats.

The Internet allows the integration of these separate groups in the same way that it has provided access to commercial products. This takes us from a few individuals fighting alone in isolated areas to having broad resources and powerful tools available.

Currently web sites are struggling to establish themselves as tool for making available biological information, weather information, satellite information, and others.

GODM will take its place along side these web sites joining in fundamentally changing our ability to view and manage our planet.

LITERATURE CITED

Arzberger, P., A. Farazdel, A. Konagaya, L. Ang, L. Shimojo, and R. L. Stevens. 2004. Life Sciences and Cyberinfrastructure: Dual and Interacting Revolutions that will Drive Future Science. New Generation Computing **22**:97-110.

BabbleFish. 2006. Babel Fish at AltaVista. babelfish.altavista.com.

Barnett, D. T., T. J. Stohlgren, C. S. Jarnevich, J. A. Ericson, T. Davern, and S. Simonson. 2006. The art and science of weed mapping. Environmental Monitoring and Assessment **(In press)**.

Bowker, G. C. 2000. Biodiversity Datadiversity. Social Studies of Science **30**:643-683.

Cecconi, A., and M. Galanda. 2002. Adaptive Zooming in Web Cartography. Computer Graphics **21**:787-799.

Celko, J. 2005. Joe Celko's SQL for Smarties: Advanced SQL Programming. Elsevier Inc., San Francisco.

Chong, G. W., R. M. Reich, M. A. Kalkhan, and T. J. Stohlgren. 2001. New approaches for sampling and modeling native and exotic plant species richness. Western North American Naturalist **61**:328-335.

Crall, A. W., L. Meyerson, T. J. Stohlgren, C. S. Jarnevich, G. Newman, and J. J. Graham. 2006. Show Me the Numbers: What Data Currently Exist for Non-Native Species in the U.S. Frontiers in Ecology and the Environment.

Crosier, C. 2004. Synergistic methods to generate predictive models at large spatial extents and high resolution. Dissertation. Graduate Degree Program in Ecology, Colorado State University, Fort Collins.

Danielsen, F., N. D. Burgess, and A. Balmford. 2005. Monitoring matters: examining the potential of locally-based approaches. Biodiversity and Conservation **14**:2507-2542.

DesktopGARP. 2006. Informatics Biodiversity Research Center, University of Kansas. www.lifemapper.org/desktopgarp.

DiGIR. digir.sourceforge.net. Distributed Generic Information Retrieval.

Douglas, D. H., and T. K. Peuker. 1973. Algorithms for the reduction of points required to represent a digitized line or its caricature. Canadian Cartographer **10**:112-122.

Drucker, H. 2007. Developing regional invasive species watch lists: Colorado as a case study. Colorado State University, Fort Collins, CO.

ECW. 2006. Earth Resources Mapping. www.ermapper.com.

Egenhofer, M. J. 1994. Spatial SQL: A Query and Presentation Language. IEEE Transactions on Knowledge and Data Engineering **6**:86-95.

Einstein, A. 1948. A Message to Intellectuals.*in* C. Seelig, editor. Ideas and Opinions by Albert Einstein. Bonanza Books, New York.

Ellisman, M. L. 2005. Cyberinfrastructure and the Fugure of Collaborative Work. Issues in Science and Technology **Online**.

ERMapper. 2006. ER Mapper. www.ermapper.com.

ESRI. 2001. ESRI Data & Maps. Environmental Systems Research Institute, Inc.

Evenden, G. I. 1990. Cartographic Projection Procedures for the UNIX Environment - A User's Manual, Open-File Report 90-284. United States Departement of the Interior Geological Survey, Woods Hole, MA.

Feyerabend, P. 1998. How to defend society against science.*in* E. D. Klemke, R. Hollinger, and D. W. Rudge, editors. Introductory Readings in the Philosophy of Science. Prometheus Books, Amherst, New York.

Foster, I. 2005. Service-Oriented Science. Science **308**:814-817.

Fox, E. A., M. A. Goncalves, M. Luo, Y. X. Chen, A. Krowne, B. P. Zhang, K. McDevitt, M. Perez-Quinones, R. Richardson, and L. N. Cassel. 2004. Harvesting: Broadening the field of distributed information retrieval. Distributed Multimedia Information Retrieval **2924**:1-20.

GBIF. 2006. Global Biodiversity Information Facility. www.gbif.org.

GDAL. 2006. Geospatial Data Abstraction Library. www.remotesensing.org/gdal.

GEON. 2006. Geosciences Network www.geongrid.org.

Gerner, S. J. 1995. Chaos and The Evolving Ecological Universe. Gordon and Breach, Amsterdam.

GISIN. 2006. Global Invasive Species Information Network. www.gisinetwork.org.

Glassner, A. S. 1990. An Introduction to Ray Tracing. Academic Press.

Globe. 2006. The Globe Program. www.globe.gov.

GoogleEarth. 2006. Google Inc. earth.google.com.

Graham, J. J. 2006. Chapter 1. A Global Organism Detection and Monitoring System for Non-Native Species. Department of Forestry Watershed and Rangeland Stewardship, Colorado State University, Fort Collins, Colorado, USA.

Green, J. L., A. Hastings, P. Arzberger, F. J. Ayala, K. L. Cottingham, K. Cuddington, F. Davis, J. A. Dunne, M.-J. Fortin, L. Gerber, and M. Neubert. 2005. Complexity in Ecology and Conservation: Mathematical, Statistical, and Computational Challenges. BioScience **55**:501-510.

Halpin, P. N., A. J. Read, B. D. Best, K. D. Hyrenbach, E. Fujioka, M. S. Coyne, L. B. Crowder, S. A. Freeman, and C. Spoerri. 2006. OBIS-SEAMAP: developing a biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles. Marine Ecology Progress Series **316**:239-246.

JPL. 2006. Jet Propultion Laboratory. www.jpl.nava.gov.

Kamath, Y. H., R. E. Smilan, and J. G. Smith. 1993. Reaping benefits with object-oriented technology. AT&T Technology **72**:14-24.

Kuhn, T. S. 1962. The Structure of Scientific Revolutions. Routledge, New York

Lepczyk, C. A. 2005. Integrating published data and citizen science to describe bird diversity across a landscape. Journal of Applied Ecology **42**:672-677.

Longley , P. A., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. 2001. Geographic Information System and Science. John Wiley & Sons, Ltd., Chichester, England.

Longley, P. A., M. F. Goodchild, D. J. Maguire, and D. W. Rhind. 2001. Geographic Information System and Science. John Wiley & Sons, Ltd., Chichester, West Sussex, England.

Mack, R. N., D. Simberloff, W. M. Lonsdale, H. Evans, M. Clout, and F. A. Bazzaz. 2000. Biotic invasions: Causes, epidemiology, global consequences, and control. Ecological Applications **10**:689-710.

Magnuson, J. J., and C. J. Bowser. 1990. A Network for Long-Term Ecological Research in the United-States. Freshwater Biology **23**:137-143.

MapQuest. 2006. MapQuest, Inc. www.mapquest.com.

Maurer, S. M., R. B. Firestone, and C. R. Scriver. 2000. Science's neglected legacy. Nature **405**:117-120.

Morisette, J. T., C. S. Jarnevich, A. Ullah, W. Cai, J. A. Pedelty, J. E. Gentle, T. J. Stohlgren, and J. L. Schnase. 2006a. A tamarisk habitat suitability map for the continental United States. Frontiers in Ecology and the Environment **4**:11-17.

Morisette, J. T., C. S. Jarnevich, A. Ullah, W. J. Cai, J. A. Pedelty, J. E. Gentle, T. J. Stohlgren, and J. L. Schnase. 2006b. A tamarisk habitat suitability map for the continental United States. Frontiers in Ecology and the Environment **4**:11-17.

NDFD. 2006. National Digital Forcast Database. www.weather.gov/ndfd.

NEON. Networking and Informatics Baseline Design. NEON Project Office.

NEON. 2006. National Ecological Observatory Network. www.neoninc.org.

Nieumwenhuijs, S. 1995. Oracle Multidimension: new frontiers in spatial data management. GIS Europe: Europe's Geographical Information System Magainze **4**:40-42.

OpenModeller. 2006. openmodeller.sourceforge.net.

Peuquet, D. J. 1994. It's about Time: A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems. Annals of the association of American Geographers **84**:441-461.

Pilevar, A. H., and M. Sukumar. 2005. GCHL: A grid-clustering algorithm for high-dimentional very large spatial data bases. Pattern Recognition Letters **26**:999-1010.

Pimentel, D., L. Lach, R. Zuniga, and D. Morrison. 2000. Environmental and economic costs of nonindigenous speces in the Unites States. BioScience **50**:53-65.

Prasher, S., X. Zhou, and M. Kitsuregawa. 2003. Dynamic Multi-Resolution Spatial Object Derivation for Mobile and WWW Applications. World Wide Web: Internet and Web Information Systems **6**:305-325.

Pun-Cheng, L. S. C., L. Zhilin, and W. Gao. 2004. An Automated System for Multi-Scale Vegetation Mapping. Cartographica **39**:89-95.

Rejmánek, M., and M. J. Pitcairn. 2002. When is eradication of exotic pest plants a realistic goal?*in* V. C.R. and C. M.N., editors. Turning the tide: the eradication of invasive species. IUCN SSC Invasive Species Specialist Group. IUCN, Gland, Switzerland.

Robinson, A. H., J. L. Morrison, P. C. Muehrcke, A. J. Kimerling, and S. C. Guptill. 1995. Elements of Cartography. John Wiley & Sons, Inc., Hoboken, New Jersey.

Schnase, J. L., J. Cushing, M. Frame, A. Forndorf, E. Landis, D. Maier, and A. Silberschatz. 2003. Information technology challenges of biodiversity and ecosystems informatics. Information Systems **28**:339-345.

Stefanovic, N., J. Han, and K. Koperski. 2000. Object-Based Selective Materialization for Efficient Implementation of Spatial Data Cubes. IEEE Transactions on Knowledge and Data Engineering **12**:938-958.

Stohlgren, T. J., D. Barnett, C. Flather, P. Fuller, B. Peterjohn, J. Kartesz, and L. L. Master. 2006. Species richness and patterns of invasion in plants, birds, and fishes in the United States. Biological Invasions **8**:427-447.

Stohlgren, T. J., and J. L. Schnase. 2006. Risk analysis for biological hazards: What we need to know about invasive species. Risk Analysis **26**:163-173.

TNC. 2006. The Nature Conservancy. tncweeds.ucdavis.edu/control.html.

UDDI. 2006. Universal Description, Discovery and Integration. www.uddi.org.

VegBank. 2006. Ecological Society of America's Panel on Vegetation Classification. www.vegbank.org.

Wang, W., J. Yang, and R. Muntz. 1997. STING: A Statistical information grid approach to spatial data mining.*in* Proceedings of the 23rd Very Large Databases Conference (VLDB 1997), Athens, Greece.

Wilcove, D. S., D. Rothstein, J. Dubow, A. Phillips, and E. Losos. 1998. Quantifying threats to imperiled species in the United States. BioScience **48**:607-615.

WIMS. 2006. tncweeds.ucdavis.edu/wims.html. Weed Information Management System.

WMS. 2006. Web Mapping Service. www.opengeospatial.org/standards/wms.

Yang, B. 2004. A multi-resolution model of vector map data for rapid transmission over the Internet. Comptuers and Geosciences **31**:569-578.

Zhilin, L., and A. Ho. 2004. Design of Multi-Scale and Dynamic Maps for Land Vehicle Navigation. The Cartographic Journal **41**:265-270.

Zhou, M., and M. Bertolotto. 2005. Efficiently Generating Multiple Representations for Web Mapping. 5th International Workshop, W2GIS. Springer-Verlag, Lausanne, Switzerland.

Zhou, X., S. Prasher, S. Sun, and K. Xu. 2004. Multiresolution Spatial Databases: Making Web-based Spatial Applications Faster. Pages 36-47 *in* Proceedings of APWeb 2004 (invited paper) (LNCS 3007), Hangzhou, China.