Collecting and Using Geospatial Data in the Field: An Extensible Framework and Testbed

Sarah M. Nusser, Leslie L. Miller, George F. Covert
Iowa State University

Michael F. Goodchild, Keith C. Clarke
University of California, Santa Barbara

Project Summary

A wide array of federal, state, and local agencies collect and analyze data to carry out their missions, evaluate programs and services, or provide information on characteristics of the population and the economy. A significant portion of these activities require mobile data collection in the field, although the nature of the field data collection environment is extremely variable across application areas. Although data gathering efforts have traditionally focused on capturing text and numeric data, spatial information may play an important role in the planning, navigation, data collection (as an object of observation or as context), editing, and analysis phases of the data collection process.

We believe that a new paradigm is needed for computer-assisted data collection that broadens the types of data used as critical information resources and as objects of data collection. Towards that end, we propose to conceive, develop, and test an extensible framework to support the collection and use of geospatial data in the field. The proposed activities are designed to meet five key objectives:

1. Develop a model documenting and formalizing the infrastructure, tools, and key capabilities required to support a flexible and extensible field data collection system.
2. Conduct research on computer science tools and associated information technologies required to fully integrate digital geospatial data into the data collection process.
3. Conduct research on infrastructure components that are needed to implement the system in a manner that limits the complexity of the system from the vantage point of the user in the field.
4. Investigate emerging field data collection technologies to determine how the usage of geospatial data is transformed by these new interfaces.
5. Explore the framework model and research developments in an application environment by developing prototype components and testbeds that correspond to agency data collection settings.

We have identified six tasks to address the research objectives. Briefly, we intend to work with representatives from government, private industry, and related academic efforts to develop:

- a user-driven framework model,
- a conceptual framework for conflation of heterogeneous geospatial data for field use,
- a multi-agent system to support tools required to use and collect geospatial data in the field,
- interoperable searching and discovery mechanisms for prepared, existing, and potentially unknown sources of data,
- object-oriented warehouse designs for the field data collection environment, and
- evaluations of emerging field technologies and their impact on user activities.

Underlying these objectives and tasks is our desire to develop a framework that generalizes as much as possible to field data collection across a range of government and private applications. While these components are described separately, it is important to recognize that there is considerable integration among them and that there are substantial benefits to studying them together, within an integrated research project. The interrelationships will be most obviously expressed in the model and testbed development.

The model framework, research activities, and testbed environments will be pursued by establishing and interacting with interdisciplinary teams of advisors and technical resources from government agencies, academia, and industry. We intend to work closely with agencies and private vendors to ensure that technologies developed as part of the research program are relevant to specific application areas and made available as commercial products. We expect

that research results will be broadly applicable in non-mobile data collection environments, as well as for uses of digital information resources by other constituencies, including the public.

The investigators at Iowa State University and University of California, Santa Barbara have substantial experience in large-scale survey data collection efforts that utilize complex information resources across a national infrastructure involving handheld and desktop computers; object-oriented databases, view systems for handling heterogeneous data sources, mobile agent systems for processing text collections and populating data warehouses, and intelligent agents for biological applications; and the geospatial context, through the Alexandria Digital Library and the National Center for Geographic Information and Analysis (NCGIA) in problems of integrating geospatial data from disaggregated sources, and mobile computing technologies.

Collecting and Using Geospatial Data in the Field: An Extensible Framework and Testbed

Project Description

1 Introduction

A wide array of federal, state, and local agencies collect and analyze data to carry out their missions, evaluate programs and services, or provide information on characteristics of the population and the economy. A significant portion of these activities require mobile data collection in the field, although the nature of the field data collection environment is extremely variable across application areas. For example, sample surveys are conducted using highly structured protocols, disaster damage reporting involves unpredictable conditions and the need to rapidly record ad hoc observations, and crime scene documentation requires a mixture of protocol-driven and adaptive data collection. The scope of a data collection effort may involve a small number of locations related to a crime scene or disaster area, or large numbers of field observations on hundreds of thousands of households or parcels of land for a sample survey.

Although data gathering efforts have traditionally focused on capturing text and numeric data, data collection and related processes commonly include a geospatial component. Geospatial information may play several roles in the data collection process:

* in navigating to a location associated with the object of data collection (e.g., household, field, plot, crime scene, disaster area),
* as an information resource providing auxiliary information needed by the officer or recorder (e.g., medical resources), or as context for interpreting conditions observed at a location (e.g., landscape features that aid in interpreting damage or existing land cover or use),
* as the object of data collection *per se* (e.g., delineating the boundaries of a farm field or other geographical feature), or
* in providing thematic information for sampling or analyses (e.g., as an auxiliary variable to determine selection probabilities, or to form an independent variable in regression analyses).

We believe that a new paradigm is needed for computer-assisted data collection systems that broadens the types of data used as critical information resources and as objects of data collection, in particular to include geospatial data in a flexible and extensible way. In establishing a new framework, we assume that collection of data will require the presence of human intelligence, and that the field data gatherer requires information services in support of decisions being made during the collection process. We are particularly interested in those decisions that benefit from the use of geospatial data. Such data may be obtained through direct measurement (e.g., global positioning systems or annotation of images), use of various sensors (e.g., pilotless aircraft), search and retrieval of data from digital libraries, and downloading prepared data from special stores. In the latter case, additional services will be required to support the preparation of data. All of these services will be subject to the limitations of field technology: bandwidth and reliability of communications links, size and capacity of field processing and storage technology, and limitations on interaction modalities (rugged devices, no keyboards, etc.). The principles we develop must be invariant across changes in technology, and hence poised to exploit new field capabilities as they become practical and broadly affordable. We will assume from the outset that many geospatial applications will require all three spatial dimensions, and also time, and avoid limiting our research to the two-dimensional, static case.

Current Information Technologies

The use of paper field data collection forms has declined over the last decade, although this format remains an important option in many survey applications (e.g., in farms surveys requiring extensive mobility). When computer-assisted data collection tools are employed, the most common approach is to use a laptop or data recorder in the field, and then upload collected data to a larger computing system in an office environment (see e.g., Gilluly, 1994; Nichols and Appel, 1994; Shanks, 1994; Hansen, 1996; Forster and Snow, 1995; Couper *et al.*, 1998). More recently, systems have been developed using handheld computer clients that communicate with a GPS unit and with a central database client-server system (Nusser *et al.*, 1996; Nusser and Thompson, 1997). Other researchers are beginning to explore the potential of wearable computers and augmented reality in a mobile environment, as these technologies become practical. For example, the need to access blueprints while working on complex assemblies has already driven the development of special eyeglasses that allow workers to view an assembly and a blueprint simultaneously; this technology has obvious applications in the field, where it can be used to provide access to maps, results of earlier field work, or even simulations of the impacts of change.

Software tools have been developed to record, store, disseminate, and analyze text and numeric data using client-server technologies. In survey sampling, software such as BLAISE and CASES are used to collect and store respondent data, but do not incorporate geospatial images and cannot be ported to the small platforms required in highly mobile situations. FieldWorker, Baker GeoResearch, and PenMap all provide geospatial data collection facilities, but they are not effective in large-scale structured survey environments. None of these products take full advantage of emergent object-oriented and integrated information system technologies or of information resources presently available via the Web, such as digital geospatial libraries and other data repositories.

Commercial database systems products (e.g., Spatial Data Blades, Oracle’s Spatial Data Option, ESRI’s Spatial Data Engine) that serve geographic information systems (GIS's) by storing geospatial objects are available for the needs of desktop environments. Some examples of utilities and pilot projects that provide access to geospatial data include the Alexandria Digital Library (alexandria.sdc.ucsb.edu), Census Bureau Tiger Map Server (tiger.census.gov/cgi-bin/mapsurfer), Massachusetts Institute of Technology’s orthophoto server (ortho.mit.edu) and Microsoft’s TerraServer (terraserver.microsoft.com), and the U.S. Geological Survey’s map service (edcwww.cr.usgs.gov/doc/edchome/ndcdb/ndcdb.html). However, the requirements for field data collection are radically different from specifications that drive existing GIS's and digital libraries.

Most of today’s GIS architectures were designed for the office environment, assuming desktop workstations or PCs with persistent network connections. Without substantial rethinking, such architectures are not able to integrate or otherwise exploit the rich range of uses of geospatial data that are essential in the field or on the street. Perhaps serendipitously, key enabling technologies—particularly the Web, global positioning systems (GPS), cellular and other wireless communications technologies, as well as low-cost handheld devices—are ripe for integration into powerful, portable geospatial information systems. Such systems are the next generation of field data collection and the focus of this research initiative. Just as important, such rapid technological change is likely to continue for a number of years; developing flexible approaches to incorporating such improvements rapidly must be a priority in federal field data collection initiatives.

Emergent Technologies

Recent technological developments make it possible for us to consider a new form of computing that is *itinerant*, *distributed*, and *ubiquitous* (IDU), characterized by devices that support computing anywhere, at any time. Reductions in size and weight, and improvements in battery technology, now allow computers to move with people, rather than tying people to offices and desks. Wireless communication technologies, such as CDPD (Cellular Digital Packet Data) and the satellite-based Iridium system, support a range of modes of interaction independent of traditional ties to the fixed locations of copper wire and fiber. New portable devices are now approaching the ruggedness and specialized modalities required for field use. On the horizon are systems that integrate voice, haptic, and head and eye tracking modalities as alternative methods of user input and extraction.

Infrastructure components are also rapidly advancing to provide the data richness to support unstructured data (i.e., text, images, etc.). For example, the object-oriented paradigm is increasingly adopted as the basis for designing information systems. In addition to providing the ability to process complex data types, the object-oriented paradigm facilitates a modular system design, thereby increasing the level of code reuse in complex and dynamic applications. The use of data warehouses is also rapidly expanding. Storing both the tools and data in the same system has proven to be very valuable for both businesses and scientific applications. Mobile agents are generating new opportunities as well. Mobile agent systems augment the benefits of the client-server model with flexibility in computational load balancing and reductions in information transferred over the network. The security concerns generated by this flexibility are currently being addressed. When supplemented by more traditional connections (e.g., CORBA), mobile agent systems can provide a safe and effective means of rapidly developing complex distributed systems. Finally, mediators have proven extremely useful in systems designed to integrate heterogeneous data sources. In the application areas relevant to the proposed research, metadata has been used to create rule bases to help locate the required data and to bridge the semantic differences.

A New Framework for Computer-Assisted Data Collection in the Field

To fully integrate geospatial data and other information resources to support field data collection, we believe that a new paradigm for computer-assisted field data collection systems is needed. This framework must be designed to:

- support effective use of digital geospatial information resources and other data stores in a field environment,
- broaden the concept of computer-assisted field data collection beyond recording data to include searching and discovery of distributed information resources,
- be effective in the full range of field data collection environments, recognizing technical and environmental factors, as well as application requirements, and
- take full advantage of current and emerging integrated information system technologies.

The broad goal of this proposal is to develop and prototype an extensible framework and related infrastructure components to support the effective use and collection of geospatial data in a distributed and mobile field environment.

This initiative grew out of work at Iowa State University over the last four years in handheld computer-assisted survey information collection (CASIC) with the U.S. Department of Agriculture (USDA) Natural Resources Conservation Service (NRCS), described in more detail below. During 1999, we engaged in a process of information exchange among principal investigators, researchers in related fields, federal and state agency stakeholders, and private industry to develop a proposal for funding the development of an extensible framework that can be applied broadly to data collection applications requiring digital geospatial materials. We worked with several federal agencies (Bureau of the Census; Bureau of Justice Statistics; Bureau of Labor Statistics; Bureau of Transportation Statistics; Department of Agriculture's Forest Service, National Agricultural Statistical Service, and Natural Resources Conservation Service; Department of Defense's Air Force Research Laboratory, and National Imagery and Mapping Agency; Energy Information Agency; National Center for Health Statistics; and U.S. Geological Survey), and discussed possible opportunities with state agencies in Iowa (Information Technology Services, Transportation, Public Safety, and the Governor's Traffic Safety Bureau). We also investigated research and technology transfer opportunities with scientists in academia (Massachusetts Institute of Technology, Stanford University, University of California Santa Barbara, University of California San Diego, University of Illinois Urbana-Champaign, and University of Southern California) and private industry (Marconi, Baker GeoResearch, Condor Earth, FieldWorker, SRI International).

Based on these discussions, we developed a new conceptual model for collecting and using geospatial data in the field, established key objectives, and built a project team. This team's research strengths and interests form the basis of a carefully planned, yet aggressive, research program, including strong involvement from the interested government agencies and private organizations in both design and evaluation of the proposed prototype capabilities. The specific research initiatives include:

- developing a comprehensive model for conceptualizing computer-assisted data collection systems from an integrated information system perspective;
- conducting research on software tools for using geospatial data, supporting infrastructure components, and the use of emerging field technologies; and
- developing testbed environments to prototype, exercise, and extend the model, infrastructure components, and supporting tools.

Relevance to Research Needs and Program Initiatives

The University Consortium for Geographic Information Science (UCGIS) has identified Distributed and Mobile Computing as one of its ten Research Challenges (www.ucgis.org), and the research proposed here directly addresses many of the issues identified by UCGIS in this area. The report of a workshop sponsored by NSF in early 1999, *Geographic Information Science: Critical Issues in an Emerging Cross-Disciplinary Research Domain*, identified field research as one of three areas of work that will be most strongly influenced by research and development in the next decade. A GIS workshop conducted in July 1998 under the auspices of NSF's Digital Government program, *Toward Improved Geographic Information Services within Government*, also identified many aspects of this proposal as having significant potential to advance the objectives of the program. Finally, the proposed research objectives address research needs outlined in the National Research Council (1999) report *Distributed Geolibraries: Spatial Information Resources.*

The research outlined in this proposal is of substantial interest to government and private industry, as indicated in the attached supporting letters. We expect to work closely with representatives from these organizations, ensuring that research initiatives and testbed development contribute to applications of interest to these groups and to the development of a more comprehensive view expressed via the framework models. This research will support the educational objectives of NSF by funding the disciplinary and multidisciplinary activities of graduate research assistants at Iowa State University and University of California, Santa Barbara, and by providing opportunities to government and private industry staff to participate in the research.

In the following sections, we present background information on current approaches to computer-assisted data collection and the challenges these systems present for future development. The specific objectives of the proposed research are then outlined and discussed. Finally, a comprehensive work plan and resources needed to conduct the proposed research are presented.

2 Current Approaches to Computer-Assisted Data Collection

Overview

In the 1970s, computer-assisted data collection systems were developed for sample survey applications (referred to as computer-assisted survey information collection, or CASIC, systems). In its original conceptualization, CASIC software was intended to provide a tool for key entry and questionnaire flow control during the telephone interviewing process, using a suite of stand-alone desktop PCs in an office environment. As computer technologies evolved, CASIC system design migrated to a client-server model in the office environment, and laptops enabled computer-assisted face-to-face interviews in households. More recently, handheld computers have been utilized to screen households on the doorstep of a housing unit (Rachel Carson, Research Triangle Institute, personal communication), conduct surveys supporting the Consumer Price Index (Bosley *et al*., 1998), and facilitate large scale natural resource surveys using a client-server system (Nusser *et al.*, 1996; Nusser and Thompson, 1997). In each of these applications, analog geospatial data are part of the survey process (paper maps and written directions are carried), but are not explicitly incorporated into the CASIC system as digital geospatial data.

Computer-assisted data collection for less structured environments involving geospatial data collection has been developed mainly in private vendor applications, as noted earlier. However, these products utilize current and emerging information technologies in limited ways, with inadequate editing and completion control for scientific data collection purposes. Without further research along the lines proposed here, such efforts will remain of limited value to the federal data collection community.

An Example of Current CASIC Systems

As an example of an advanced computer-assisted data collection system, consider the CASIC system developed by the USDA Natural Resources Conservation Service and Iowa State University to conduct the National Resources Inventory (NRI) in mobile office and field settings (Nusser and Goebel, 1997). The NRI sample consists of approximately 300,000 area segments (typically 65 ha) throughout the U.S. Data are collected via remote sensing (using low altitude photos, maps, field office record abstraction) and in some cases, field observations. Variables recorded for each sample area segment include the surface area for land categories (e.g., water, federal, farmsteads) and length measurements for water bodies and streams. Most of the important subject matter variables are collected on one to three randomly selected sample points within the area segment; point variables include land cover/use, soils information, agricultural practices, forest type, and wetland classification. Over 200 collection teams distributed throughout the U.S. are involved in a full survey effort, with supporting conditions such as computing and physical data collection environments varying widely.

A client-server architecture was developed in 1996 to support handheld computer-assisted data collection. The system consisted of:

* redundant central database servers with RAID storage,
* an Oracle database containing tracking variables and historical and newly collected data for each area segment,
* redundant front-end servers to negotiate database requests,
* about 500 handheld computers (Apple Newtons) equipped with a computer-assisted data collection form for recording values for each variable, and
* TCP/IP-based communication via a variety of modes (wireline, wireless, local area network, and Internet) between the front-end servers and the handheld client.

A data gatherer logs on to a front-end server for a short period to query the database and request samples to be worked on. During data collection, the client software notifies the user of apparent inconsistencies among collected variables, as well as with historical data stored onboard. These problems are resolved and the data gatherer returns the sample data to the server. Except for interfaces used to define queries and samples to be returned, most of the negotiation between the handheld computer and server system is hidden from the data gatherer. Security is maintained using UNIX system user authentication and user access restrictions for the servers and in the database corresponding to geographically assigned workloads.

Challenges in Current CASIC System Design

The NRI CASIC system has provided significant improvements over the former stand-alone desktop systems. It has introduced small and large-scale mobility into the data collection environment, has improved data quality by effectively incorporating editing into the data gathering process, and has greatly reduced the time required to resolve problems during post-data collection processing, when compared with past efforts. However, a number of limitations are presented by this paradigm, as noted below.

- Several components of this survey are explicitly geospatial, yet there is no mechanism to utilize digital geospatial data in the field. Currently, a mixture of analog and digital materials is used to collect data and it is difficult to integrate these sources in their current forms. The quality of data gathered and the efficiency of the various components of the process can be greatly improved by developing a framework that facilitates integration of digital geospatial data into these processes.
- Structured relational data formats for input, storage, and queries have provided the utilities needed to collect data according to prescribed scientific protocols. However, the use of these tools requires special code to be developed to accomplish these functions. They lack the flexibility to incorporate geographic data and do not take advantage of integrated information system tools.
- The client-server model requires an open connection in order to exchange information, limiting transfers to relatively small requests. Software agents could be used to send complex queries or request analyses that exceed the shorter connection times appropriate for the current system, returning results of the request at a later time.
- The current system is designed for a single purpose. Hardware and some code are reused as new surveys are deployed. However, the system is not inherently modular or readily adaptable for new data collection efforts or interaction with activities in other parts of the agency or in other agencies.

While the client-server CASIC model has served the data collection process well, to fully exploit the potential of digital geospatial data and emerging information technologies, a new paradigm is needed for mobile field data collection.

3 Objectives

We propose to conceive, develop, and test an extensible framework to support the collection and use of geospatial data in the field. To achieve this goal, we will meet five key objectives:

6. Develop a model documenting and formalizing the infrastructure, tools, and key capabilities required to support a flexible and extensible field data collection system.
7. Conduct research on computer science tools and associated information technologies required to fully integrate digital geospatial data into the data collection process.
8. Conduct research on infrastructure components that are needed to implement the system in a manner that limits the complexity of the system from the vantage point of the user in the field.
9. Investigate emerging field data collection technologies (e.g., wearable computers, augmented reality environments such as smart glasses, and new modalities such as voice, haptic, and eye tracking) to determine how the usage of geospatial data is transformed by these new interfaces.
10. Explore the framework model and research developments in an application environment by developing prototype components and testbeds that correspond to agency data collection settings.

Underlying these objectives is our desire to develop a framework that generalizes as much as possible to field data collection across a range of federal, state, and local agencies, as well as citizen applications that benefit from use of digital geospatial data. In the following section, we discuss our approach to achieving each of these objectives.

4 Proposed Research

In this section, we provide an intellectual framework for the five proposed research areas. While these components are described separately, it is important to recognize that there is considerable integration among them and that there are substantial benefits to studying them together, within an integrated research project. These relationships are most obvious in the framework model and will be directly expressed via testbed development. Further elaboration of the relationships among specific tasks for achieving these goals is provided below and via the work plan in Section 5. This sort of integration and overlap is necessary to properly advance the complexities of the problems posed by using geospatial data during field data collection, and is the basis for forming interdisciplinary teams. Although the five research areas are specifically targeted to the current problem, they are also broadly applicable in non-mobile data collection environments, as well as for other uses of digital information resources by government, industry, academia, and the public at-large.

Objective 1: Conceptual Model for Framework

A critical research objective is to develop a model documenting and formalizing the infrastructure, tools, and key capabilities required to support a flexible and extensible field data collection system. The model will be designed to incorporate a broad suite of digital data formats and information resources, and will support future development of the infrastructure as new uses and technologies evolve. This model forms the basis for guiding the research program and testbed development, and will be adapted in response to findings discovered through research and testbed activities, as well as through feedback provided by the participating agencies.

In considering the types of interactions and environments that must be planned for, we envision a triangular model (Fig. 1). The nodes represent three different components, each of which has unique characteristics corresponding to the data collection perspective (data user or producer), physical environment, information resources, computing capacity, access, and customization control. More specifically, the three components are:

* the field worker, who seeks information to support the data collection process and returns data resulting from collection activities,
* the organizations sponsoring and/or involved in data collection, which support special data preparations, proprietary stores, and repositories for collected information, and
* other information resources accessible through digital libraries or from organizations not specifically collaborating with the data collection effort.

The central portion of the triangle represents the infrastructure required to support information exchange and discovery among these actors, resources, and environments. Fig. 2 presents an initial infrastructure model that schematically describes the pathways of this exchange. The field data collector is connected via traditional (e.g., hard coded) wireline or wireless connections to servers that are configured to support data collection (solid lines in Fig. 2). In some cases, this takes the form of logging directly into a specially designed data repository (upper DCR box in Fig. 2), in a manner similar to current CASIC models. Alternatively, the field worker may log in to a local data warehouse. The warehouse server is designed to receive queries, search repositories, and prepare data or data views for the field worker's use, drawing from internal databases created specifically for the survey (upper DCR boxes in Fig. 2) and from other information resources available via the Internet (OIR boxes in Fig. 2). The interaction between the local data warehouse and multiple DCR and OIR data sources is negotiated via mediators, agents, and views (dashed lines in Fig. 2), whose complexities are hidden from the field user.

We propose to further develop these conceptual models, addressing such issues as:

* the various categories of views required for unique classes of field users with specific input and output devices, environmental constraints, and data collection applications;
* the protocols, services, and data views required within participating data collection organizations hosting the DCRs; as well as
* rules guiding infrastructure choices when supporting the use of data from external digital information resources (OIRs).

As noted earlier, these models will be developed in parallel with the specific research areas and testbed activities.

Objective 2: Tools for Field Use of Geospatial Data

We will conduct research on information technologies required to fully integrate digital geospatial data into the data collection process. Although great progress has been made in developing geographic information systems over the

Figure 1. Model for field data collection environment in relation different data collection perspectives, physical and computing environments, and types of information resources.

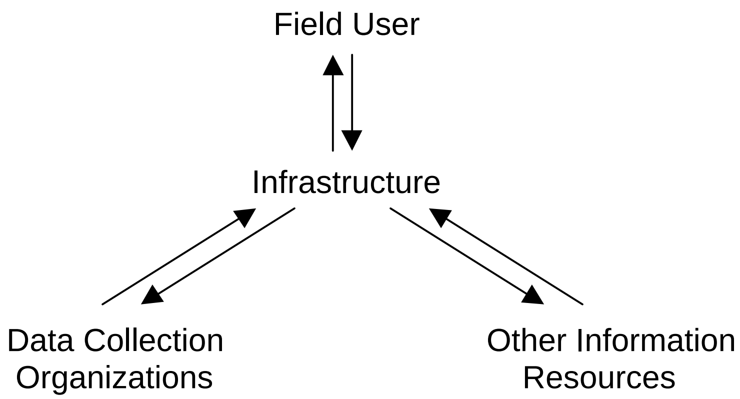
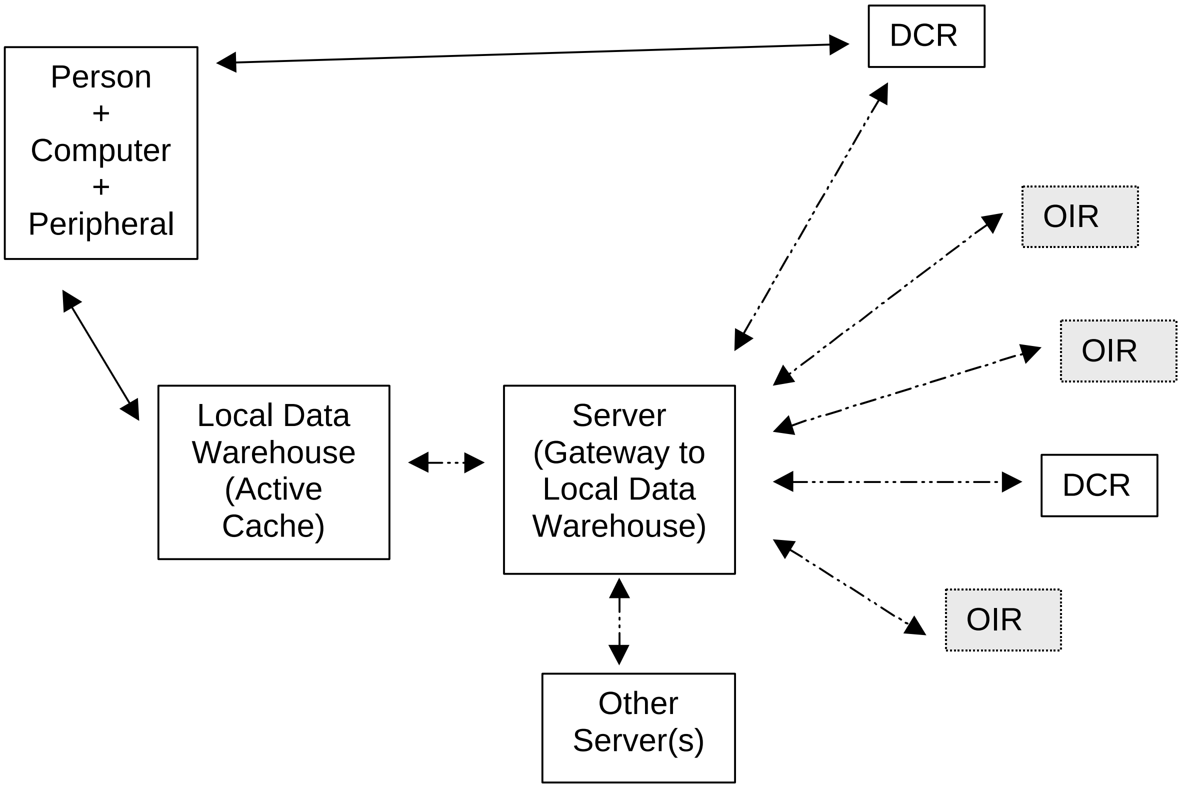


Figure 2. Infrastructure model. Dashed lines imply mediated information exchange or knowledge acquisition. Solid lines imply traditional (e.g., hard coded) information exchange. DCR (unshaded box) represents a repository hosted by one of the data collection organizations, OIR (shaded box) represents other digital information resources hosted by organizations not involved in data collection.



past three decades (for a recent review see Longley *et al.*, 1999), little of this effort has gone into researching the unique requirements of field data collection, or into developing relevant technologies. GIS has traditionally functioned as a stand-alone application, focused on analysis of existing data, and since the mid 1980s has been supported on the desktop. More recently it has become common for GIS data to be downloaded from archives using the WWW and FTP, and a number of significant archives of geospatial data now exist. But field data collection differs from the traditional GIS environment in many significant ways.

Information technology is used in support of the creation of information, rather than the analysis of existing information, raising questions about how one represents partial knowledge, measures improvements in the state of knowledge, or knows when knowledge is complete. We will develop data models for partial knowledge, based on adaptation of existing GIS data models (for reviews of GIS data models see Longley *et al.*, 1999; Molenaar, 1998), and associated metrics of completeness. Such extensions to existing GIS data models have been identified as a significant component of the research agenda in geographic information science (UCGIS, 1996).

Improvements in communications technology and in digital libraries offer the prospect that field data collection will be able to rely on real-time access to extensive stores of information in distributed libraries that are searchable using geographic location as the primary key (National Research Council, 1999). To date, little attention has been paid to issues of field access to digital libraries. One exception is the Stanford University digital library group, which plans to undertake such research in collaboration with University of California, Santa Barbara's Alexandria Digital Library (ADL) under the new Interlib consortium, and has already experimented with field access in the context of electronic commerce (Daswani and Boneh, 1999). We have been unable to find any previous work on how to access such information to meet field needs and under the constraints of field operation. We will work with ADL and the larger Interlib consortium (UC Santa Barbara, UC Berkeley, Stanford, the California Digital Library, and the San Diego Supercomputer Center) to adapt existing clients to the requirements of field operation. We will integrate this work with the efforts of major database vendors (Oracle, Informix), and explore the use of spatial database engines (ESRI's SDE, Oracle's Spatial Data Option). ADL currently uses Informix as its underlying database.

The ability to access multiple sources of information for the same area raises important issues concerning interoperability (Goodchild *et al.*, 1999), visualization, inaccuracies in positioning, and other conflicts. We use the term *conflation* here to refer to this suite of problems (the term *fusion* is commonly used for the conflation of imagery; (e.g., Benediktsson and Landgrebe, 1999). For example, a field client accessing a digital orthophoto and a vector representation of the area's surficial hydrology will have to resolve such issues as how to view them simultaneously; how to remove any obvious positional discrepancies between them; and how to deal with different projections and Earth datums. We propose to develop a comprehensive approach to conflation, addressing all types of geospatial data. Our approach will utilize the strengths of ISU in statistics, especially spatial statistics, and of UCSB in geospatial data, and will build on prior work on limited facets of the conflation problem (e.g., Kyriakidis *et al.*, 1999; Goodchild *et al.*, 1995). We will assess the limited tools currently available from software vendors, integrate them and current literature into our general framework, and develop associated tools for field use. Because of the growing importance of this field to many applications of geospatial data, the dearth of research to date, and the scattered nature of research across many disciplines, we propose to organize an international conference on conflation in the second year of the project, providing support for young scholars and students to attend (see UCSB budget justification).

Finally, the ability to access and analyze data in the field raises interesting issues concerning real-time control of sampling. Traditionally, sampling designs have been laid out in advance of field campaigns, and remain fixed irrespective of the knowledge that is acquired in the field. By analyzing data as it is collected it may be possible to make better decisions about the need for further samples, and designs for selecting additional sample units, by developing improved methods of adaptive sampling (Thompson and Seber, 1996). We propose to develop tools for continuous monitoring and adaptation of geospatial sampling designs in the field, utilizing ancillary variables downloaded into the field, and analysis of data already collected. For example, we will investigate the implementation of methods of geostatistics (e.g., kriging methods, as described in Cressie, 1993; Burrough and McDonnell, 1998; Deutsch and Journel, 1998) in field devices to monitor and modify sampling strategies as knowledge is acquired, utilizing ancillary covariates downloaded from digital libraries. Based on analysis of data already acquired and ancillary data delivered to the field site, such tools will be capable of supporting adaptive sampling, calculating inclusion probabilities with higher selection rates for sites of interest, or suggesting modifications to linear transect sampling. The technologies that support the field worker in Fig. 1 will include agents and other mechanisms for finding and retrieving data from information resources, and tools for working with such data locally. Thus all of our research areas form a tightly integrated whole, with the prioritized tasks listed below in our work plan as key issues to be addressed by the team during the three years of planned effort.

Objective 3: Extensible Infrastructure for Field Data Collection

The third research objective seeks to design and prototype the infrastructure necessary to provide the field data collector with the ability to incorporate geospatial data into the field environment. Information retrieval and discovery from heterogeneous geospatial data sources residing on multiple hardware platforms and operating systems at different geographical locations require a robust and flexible framework for information exchange between the various data sources and clients. One of our key design principles is that users must have a simple interface to information resources, whether they are connected to a single source (e.g., data collection agency) or have the ability to interact with a wide range of data sources in an ad hoc fashion.

To provide the infrastructure for the models shown in Figs. 1 and 2, we propose to use views and mediators to locate data sources and bridge semantic differences, local object-oriented data warehouses to provide temporary storage and analysis of geospatial data, and a multi-agent system to connect the user to the data sources and support knowledge discovery and information retrieval. The three areas are examined in more detail below.

Processing heterogeneous data sources: multidatabase and mediator-based systems

Numerous changes need to be addressed to provide integrated access to heterogeneous data sources. Research on these problems has taken two directions: *multidatabase systems* (Bright *et al*., 1992; Hurson *et al*., 1994; Barsalou and Gangopadhyay, 1992; Su *et al*., 1993; Owrang and Miller, 1988;), and *mediators* (Wiederhold, 1992; Wiederhold and Genesereth, 1997; Wiederhold, 1999). The multidatabase approach hinges on traditional database techniques, while mediators use a set of rules to locate data sources and bridge semantics. At present, there is no implemented system that offers the full range of functionality envisioned by Wiederhold and Genesereth (1997). Examples of projects that have used mediators include HERMES at the University of Maryland (Subrahmanian, *et al.*), CoBase at UCLA (Chu *et al*., 1996), use of mediators in scientific data warehouses (Critchlow *et al*, 1998), and National Industrial Information Infrastructure Protocol Common Language (NCL) at the University of Florida (Su *et al.*, 1996).

We propose to use both approaches in our infrastructure design and implementation. From the multidatabase perspective, we will use object-oriented views to provide integrated access to heterogeneous distributed databases (Yen *et al*., 1998; Miller and Yen, 1995; Yen, Miller and Pakzad, 1994; Yen, Miller and Wong, 1994). In our approach, an object-oriented view is made up of three parts: (1) a data description, (2) a data derivation section, and (3) a methods section. An example of a view in our system would be to have the data description define a map for downloading to a palmtop. The derivation section would contain the code necessary to extract the spatial window and adjust the spatial resolution given in the dataset to the resolution needed in the palmtop. In the proposed research, views will be combined with mobile agents to facilitate extracting the required data from geospatial data sources. The derivation section of a view will be used as the agent "program" to extract only the data needed. Views can be designed over multiple data sources, so agent views will be used to gather information for processing at the local data warehouse level. For example, a view can be used to gather data on street information from one source and county boundaries from another source for combining at the local warehouse level.

Researchers at the San Diego Supercomputer Center (SDSC) have successfully used mediators based on metadata in their MIX system (Ludascher *et al.*, 1999, Baru *et al.* , 1998, Gupta *et al.*, 1999). The XML metadata language is the basis for the mediators and an XML-view system. We will work closely with SDSC to investigate extension of the MIX system to our problem domain and to examine possible connections between the XML-view system and the design of our object-oriented views. An initial meeting early on in the project is planned to facilitate integration of ISU, UCSB, and SDSC research initiatives in this area.

Local data processing and caching: data warehouses

Data retrieved from remote sources for data collection will require local caching for further analysis (e.g., conflation of more than one source item). This calls for a *data warehouse* as an integral part of any system based on distributed heterogeneous geospatial data sources. Object-oriented data warehouses have started to receive attention from the research community (Widom, 1995; Miller *et al.*, 1997, Miller *et al*., 1998, Thomann and Wells 1997). A data warehouse is differentiated from a database system by its capability to store both the data and a set of tools for analysis of the data. The tool set allows the data warehouse to act as an active cache in the sense that it will allow for both temporary storage of data and local processing of the data. The proposed research seeks to explore these issues by designing, implementing, and experimenting with an object-oriented data warehouse system for geospatial data. Commercial database management systems (e.g., Oracle) will be examined as the means of supporting the data management functions of the warehouse. The object-oriented view mechanism will be used to support an object-oriented data warehouse. It will provide a rich environment for storage, retrieval, application-dependent transformation, and analysis of heterogeneous data. Design of the warehouse will require close collaboration among ISU and UCSB researchers. The views used to populate the warehouse will incorporate search information developed at UCSB (e.g., the ADL query language). The toolset used in the warehouse will include both traditional data mining tools and special tools for geospatial data (e.g., conflation tools developed under Objective 2).

We will also investigate using multiple layers of the "local" data warehouse, with an emphasis on matching the level of the warehouse with the processing capability of the field computer. Of particular interest is determining the feasibility and value of putting the warehouse closer to the field user, such as utilizing a portable computer in the field (e.g., in the vehicle) to provide efficient access to high volumes of data.

Discovery and retrieval: multi-agent systems

Identification and retrieval of relevant data is essential to the success of the proposed model. It is essential that the information discovery aspects of the environment be automated as much as possible. To achieve this we propose using a two-tier system. The basic infrastructure consists of a multi-agent system supported by more traditional connections. Two levels are needed due to current security concerns for agent systems (Chess, 1998) and questions of where processing is allowed.

As commercial agent infrastructures evolve over the next few years, it is expected that many of the security concerns will be addressed. However, traditional connections, like CORBA, will continue to be needed for data sources that require control of the amount of processing done on the sources’ computers. Mobile agent environments allow the program to be moved to the data source. Our proposal allows an agency to support only retrieval by letting the agency restrict their connection within the model to a CORBA connection. As a result, the data from the agency will be retrieved and sent to the nearest system server for processing.

For the purpose of this document, an *agent* is a computer program that performs a set of tasks on behalf of a user with some degree of autonomy (Bradshaw, 1997; Finin, 1999). In order to do this, an agent has to embody a certain amount of intelligence (e.g., the ability to choose among alternative courses of action, plan, communicate, adapt to changes in the environment, and learn from experience). An agent consists of program code, a persistent internal state, and a set of attributes (e.g., movement history, authentication keys, etc.) (Honavar, 1999). Mobile agents are agents that can move in a computer network from host to host as needed (Aridor and Oshmia, 1998).

The research focus will be to design a suite of mobile agents. Several types of agents will be prove useful in supporting the proposed model (e.g., search agents, discovery agents, and mediator agents). ISU and UCSB researchers will collaborate to create the optimal suite of agents to support geospatial data. The Open Agent Architecture developed by the SRI International will be examined as a means of supporting server and mediator functions in our model (Cohen *et al.*, 1999).

In order for agents to be useful in geospatial information retrieval and discovery tasks, agents have to be both reactive as well as proactive in gathering relevant information from dynamic data sources. Of particular interest are automated information and knowledge discovery agents that acquire potentially useful information from distributed heterogeneous digital geospatial data sources. For instance, examining multiple maps, a knowledge discovery agent could infer a relationship between the contents of the maps and some event of interest. This can be achieved by incorporating machine learning algorithms into some of the agents.

The multi-agent system will incorporate mediators to make use of metadata to help locate and define the properties of the relevant geospatial data. The object-oriented view system will be incorporated into the mobile search agents to provide the basis of transforming heterogeneous data types to provide a consistent format for an application. The multi-agent system will interface with the local data warehouses to provide for local caching and analysis of the relevant data.

Prototypes for geospatial data

The focus of the research here will be to design and implement a prototype of the multi-agent-mediator-warehouse system. A suite of agent classes will be developed to support mediation using metadata, information retrieval, and knowledge discovery. A multi-level object-oriented data warehouse will be designed and prototyped to operate on geospatial data, based on agency applications. Our object-oriented view system will be adapted in cooperation with UCSB and ISU researchers to accommodate geospatial data and integrate it into the mobile agent environment.

Objective 4: Emerging Field Technologies

Clarke (1998) has recently reviewed the range of emerging wearable and mobile technologies from the perspective of geospatial applications. He argues that in this area, technological development precedes science in making it possible to collect new kinds of information and to perform new tasks, independently of previous constraints. New field technologies will continue to emerge in the coming years, motivated by requirements and applications that may have little or nothing to do with field data collection. Significant recent developments include cellular communications, highly portable miniaturized computing systems, special purpose input and output devices such as glasses and condensed keyboards, and voice synthesis and recognition software, referred to as VISIO (voice-in, sound and images out) interfaces. The ability to identify such technological developments, recognize their potential for vastly different applications, and adapt and integrate them for effective use will be crucial. Nevertheless, many are off-the-shelf technologies, in need of assembly and integration around the tasks and environment of field computing. Under this topic, by year, we will (1) investigate existing technologies that show promise as components of an integrated field computing system; (2) identify and research the implications of the prototype design for user interface design and; (3) build a prototype system from available components, creating software when necessary to fill gaps in functionality.

Field technologies can be organized along several key and largely invariant dimensions: the degree to which they achieve the itinerant, distributed, and ubiquitous (IDU) standard; the modalities of user interaction (senses used, input and output streams invoked, constraints imposed by the local environment and the computational capabilities and bandwidth); and the limitations imposed by human cognition, including attention focus and information overload. In addition, the field imposes constraints that are not significant in other application areas. For example, technologies that will work for medical and engineering applications inside buildings may well fail in the strong daylight and poorer communications typical of the field. The hardening of devices necessary for military use may not be fully necessary, but nevertheless the systems will need to work reliably in extreme environmental conditions. We will investigate this issue in detail, as it threatens to limit the general applicability of many new devices.

Our work will incorporate existing and newly developed field technologies. For example, we will consider the prototype wearable computing systems that have been developed at Oregon State University, at MIT (which supplies plans for the custom construction of their Lizzie system), Georgia Tech, Michigan, Purdue, Boeing, and at eight defense and industrial establishments (www.cs.purdue.edu/research/cse/scipad/mobicomp.html, www.cs.uoregon.edu/research/wearables/Papers). Wearable computing has an on-line journal and is in the third year of a sequence of International Conferences. Examples of commercialized wearables are available at http://hwr.nici.kun.nl/pen-computing/wearables.html. We are aware of only one major research effort in the geosciences specifically on field computing, at the University of Kent (www.cs.ukc.ac.uk/research/infosys/mobicomp/Fieldwork/index.html), and have already established links with that group. Their focus is primarily developing educational tools for outdoor laboratories using a mobile computing environment.

Objective 5: Prototype and Testbed Development

Integrated research of the nature and magnitude of that proposed here under the Digital Government program is best conducted in an environment that combines basic research with the development of prototypes and testbeds. The latter ensure that the research team learns from the experience of implementation, and that the user community has the opportunity to provide frequent and useful feedback. We propose therefore that all areas of research in this project be integrated around the development of prototypes and testbed systems, and that a stage in all of our research tasks be devoted to working with agencies and private industry to learn more about how our systems function in real environments.

Through the proposal development process, a number of federal, state and private industry partners have expressed an interest in participating in testbed developments to facilitate technology transfer. Organizations writing supporting letters (see attachment) include the Bureau of the Census; Bureau of Labor Statistics; Department of Agriculture’s Forest Service, National Agricultural Statistics Service, and Natural Resources Conservation Service; Department of Defense’s Air Force Research Laboratory and National Mapping and Imagery Agency; U.S. Geological Survey’s National Mapping Division; Iowa Departments of Information Technology Services, Public Safety, Transportation, and Governor’s Traffic Safety Bureau; and private vendors Baker GeoResearch (GPS/GIS mapping and databases), Condor Earth Technologies, Inc., (geospatial data collection software), FieldWorker Products, Ltd., (geospatial data collection software) and Marconi (software for emerging technologies). In addition, verbal interest in our research has been expressed by the Federal Geographic Data Committee (FGDC), and SRI International.

To ensure that our work is compatible with existing and emerging standards, we will also work with the Open GIS Consortium (OGC) and its specifications; the FGDC, its metadata standards and National Spatial Data Infrastructure; and with other relevant geospatial standards such as those being developed under ISO TC 211.

5 Work Plan

Scheduled Tasks

We have identified six comprehensive tasks that are required to complete the research program described in Section 4. In identifying these tasks we have been mindful of the need to prioritize, and to achieve substantial results within a reasonable time, since it will not be possible to pursue all areas of interest and potential significance. The six tasks are:

1. Create a mechanism to incorporate user-driven design in model, research, and testbed development by establishing and interacting with interdisciplinary teams of advisors and technical resources from government agencies, academia, and industry.
2. Develop and implement a conceptual framework for conflation of heterogeneous geospatial data for field applications.
3. Develop and test a multi-agent system to support tools required to use and collect geospatial data in the field.
4. Develop and implement interoperable searching and discovery mechanisms for prepared, existing, and potentially unknown sources of data.
5. Adapt and prototype object-oriented warehouse designs for the field data collection environment.
6. Evaluate emerging field technologies and their impact on user activities, the conceptual framework, and developing research.

Table 1 relates these tasks to specific areas of pursuit in each year of the investigation, and identifies links to the five research objectives outlined in Sections 3 and 4. Each task will be performed by an interdisciplinary team; research leaders and collaborating scientists are also noted in Table 1. Because of the inherent interconnections in these topics, research teams devoted to these tasks will frequently communicate their findings with one another. The integration of these areas is expressed in the conceptual framework model, will be implemented and tested through prototype and testbed development.

Project Management

Project management will be accomplished at two levels. First, an interdisciplinary advisory board will be established to provide broad guidance to the project and review results at six-month intervals. Second, a technical steering committee of a half dozen members will be created from key investigators, agency stakeholders, and industrial partners. The steering committee will hold teleconferences monthly and meet twice per year to ensure integration of research areas and to ensure that design and development approaches are most useful for participating agencies. A web site will also be established as a common workspace to post publications, preliminary results, and tools, and to provide access to prototype systems.

Outreach

Outreach will occur via several routes. The web site used by project members will have a segment that will provide public access to project results. The public web site will be created to share proposed interfaces and project components, and to provide appropriate testbed evaluation and access. Project staff will present results at conferences and publish refereed papers in their respective scientific areas as well as in other disciplines that would benefit from this technology. Included in this form of outreach are regular presentations to interested agency staff. Investigators also plan to host conferences on specific domain areas (e.g., conflation in Year 2) as well as an interdisciplinary forum to discuss cross-cutting themes (Year 3).

Project Staff and Expertise

The investigators for the grant represent a mixture of domain areas. Sarah M. Nusser (survey statistics and data collection methodologies) will be the PI for ISU, and Michael F. Goodchild (geospatial data) will be the PI on a subcontract to UCSB. Co-principal investigators are Leslie L. Miller (computer infrastructures) and George F. Covert (testbed applications) at ISU and Keith C. Clarke (emerging field technologies) at UCSB.

The investigators at ISU have substantial experience in a specific domain area (the federally-mandated USDA NRCS Natural Resources Inventory) that exhibits many important characteristics related to the proposed research. We have learned to design, build, and support complex information resources across a national infrastructure involving handheld and desktop computers. In addition, Iowa State University computer science researchers have developed object-oriented databases, view systems for handling heterogeneous data sources, and mobile agent systems for processing text collections and populating data warehouses, and are exploring intelligent agents for biological applications.

In particular, Nusser leads the Iowa State University Statistical Laboratory Survey Section and is involved with the research and operational support for a large-scale USDA NRCS Natural Resources Inventory survey program. She has direct research and applications experience in survey statistics and computer-assisted survey information collection systems for handheld computers in mobile survey environments. Miller, a faculty member in the ISU Computer Science Department, is involved with research in developing and populating data warehouses, using agent systems to detect intrusions into network systems, and application of agents and machine learning to biological problems. Covert, Associate Director of the Computation Center, has led the Computation Center’s development of production-level central database and support systems for the distributed data collection activities for USDA natural resource surveys.

Supporting investigators at ISU have experience in infrastructure technologies and large-scale application environments. The ISU-USDA survey collaboration is supported by the ISU Computation Center, which is led by Peter Siegel, ISU’s Director of Academic Information Technologies. Vasant Honavar and Johnny Wong are members of the ISU Computer Science faculty. Wong leads a project on using agents to detect intrusions into network systems. Honavar's research is focused on machine learning algorithms and the incorporation of machine learning into agent systems. Honavar is also involved in the intrusion detection project and is active in looking at applying machine learning to biological problems.

Investigators at the University of California, Santa Barbara have extensive experience with the geospatial context, through the Alexandria Digital Library (one of six major projects funded by NSF under the first Digital Library Initiative, and with continued funding under the second initiative 1999–2004), and through the National Center for Geographic Information and Analysis (NCGIA), in problems of integrating geospatial data from disaggregated sources, and mobile computing technologies.

Goodchild was Director of the National Center for Geographic Information and Analysis from 1991 to 1996; currently, he is Chair of the NCGIA Executive Committee and Director of the Varenius project, NCGIA's project to advance geographic information science. His current research interests center on geographic information science, geospatial analysis, the future of the library, and uncertainty in geographic data. Clarke is currently Director of NCGIA at UCSB and conducts research on the use of geospatial information. Terry Smith will be a supporting investigator at UCSB. Smith leads the Alexandria Digital Library effort at UCSB, with an emphasis on supporting digital geospatial data.

Funds are requested for both ISU and UCSB to support postdoctoral associates and graduate research assistants working with investigators and participating organizations. Professional staff at ISU’s Computation Center and Statistical Laboratory and UCSB will be involved in testbed and applications development.

Taken as a whole, our research experience, technical directions, and federally-supported applications have demonstrated relevance to a broad range of geospatial data and computer science problems in government, industry, and academic research. In assembling our research team, we have focused on the need to develop an effective collaborative effort to refine models for geospatial data collection and management, to take advantage of rapidly changing computer science knowledge and technology, and to generalize the models to advance both specific application domains and geospatial data collection programs broadly.

Table 1. Work to be conducted for each year by task. Also given are proposed project teams (with task leaders in italics) and links to the primary research objectives for each task. All tasks involve model development (Objective 1) and testbed and prototype generation (Objective 5).

Task & Primary Research Focus	*Year 1*	*Year 2*	*Year 3*	*Team (Task Leader in Italics)*
User-Driven Design (Objectives 1-5)	Establish advisory board & technical team Initial project meeting with stakeholders and technical team Elicit agency, field, and researcher uses, requirements Develop technology-invariant model for field uses Develop Web site for outreach and project communication	Review conceptual designs with user groups Begin incorporating knowledge into model from testbed implementation Organize an international conference on field uses of geospatial data, to share results, help to build an international community of scholars, and focus future research efforts	Test prototypes with user groups, with special emphasis on generalizing results beyond the specific testbed applications Organize a national workshop to showcase results with vendors, agencies, and academic research community	*Nusser* All PIs Agencies, vendors
Conflation (Objective 2)	Develop conceptual framework for conflation to integrate broad suite of digital geospatial data formats Research the literature on data fusion in related areas, including image processing Develop tools for conflation within the constraints of field technologies	Develop prototypes using selected testbeds (MIT, ADL, NGDC, Tiger, USGS, TerraServer, etc.) Organize international conference on conflation, to share results, help build an international community of scholars, & focus future research efforts	Test capabilities in USDA (NRCS, NASS, USFS) setting Extend, generalize, and transfer technology to other applications	*Goodchild* Nusser (sampling and survey applications) Agencies, vendors
Searching Mechanisms (Objectives 2, 3)	Specify interoperability problem details, scope requirements Identify how metadata will be used to facilitate search & discovery, in conjunction with SDSC. Integrate with agent, warehouse components	Develop prototype field client capable of searching and accessing the data resources of ADL, MIT orthophoto server, NGDC, and other selected major resources	Test capabilities in USDA setting Extend, generalize, and transfer technology to other applications	*Miller* Goodchild (geospatial data), Smith (ADL) San Diego Supercomputer Center Agencies w/ databases

Table 1 (*continued*). Work to be conducted for each year by task. Also given are proposed project teams and links to the primary research objectives for each task. All tasks involve model development (Objective 1) and testbed and prototype generation (Objective 5).

Task & Primary Research Focus	*Year 1*	*Year 2*	*Year 3*	*Team*
Intelligent Agents (Objective 3)	Develop model for multi-agent system and other elements (e.g., CORBA, mediators) Initiate interaction with SDSC Evaluate the Open Agent Architecture for integration into the infrastructure	Develop prototypes using selected testbeds (MIT, ADL, NGDC, Tiger, USGS, TerraServer, etc.) Develop methods for information exchange capabilities with federal agencies (e.g., updating Tiger)	Test capabilities in USDA setting Extend and transfer technology to other applications	*Miller* Honavar, Wong (machine learning) Goodchild (geospatial data) Nusser, Covert, Siegel (application testbed) Agencies w/ databases
Warehouses and Views (Objective 3)	Adapt object-oriented warehouse design to requirements of field data collection environment Develop object-oriented views Initiate and plan prototyping Evaluate Oracle for providing data management support for the warehouse prototype Develop alternative architectures & associated functionalities for other settings	Select architectures Integrate w/ developing prototype of agent & search infrastructure Redesign in response to testing	Continued testing and development (minor extensions) Generalize architecture for other agency applications & research uses	*Miller* Goodchild (conflation) Nusser, Covert, Siegel (application testbed) Agencies
Field Technologies (Objective 4)	Establish relationships w/ commercial vendors & suppliers Evaluate emerging commercial technologies Develop conceptual framework for evaluating field devices (type, interface mode, capacity, ruggedness, communication, bandwidth, etc.)	Evaluate emerging commercial technologies Initiate integration with prototypes Understand how user design / technology impacts infrastructure components (flexibility, robustness, scalability) Feedback to vendors	Evaluate emerging commercial technologies Feedback to vendors Generalize conceptual framework to needs of other agencies / users	*Clarke* Nusser (application testbed) Agencies, private vendors