CLOSING REPORT FOR NCGIA RESEARCH INITIATIVE 5: ARCHITECTURE OF VERY LARGE GIS DATABASES

Terence R. Smith Andrew Frank

March 31st, 1992

1 BACKGROUND

Activity with respect to Initiative 5 of the NCGIA, titled *Architecture of Very Large GIS Databases*, commenced in July, 1989. The Initiative was officially completed in August 1992, with presentations of three papers at the Symposium on Spatial Data Handling in Charleston, South Carolina by A. Frank, M. Egenhoffer and R. Barrera; by W. Kuhn; and by T. Smith. It is to be emphasised, however, that research is currently continuing on key issues that lie at the heart of the initiative. For example, research relating to spatial database and modeling support for scientific research will continue at UCSB over the next three years, under the direction of Smith and funded by NSF; research relating to the properties of spatial data that are important for optimization of spatial storage and query execution strategies will continue at Vienna, under the direction of Frank and funded by ESPRIT. It should also be pointed out that other research projects that are likely to have major implications for this area of research are now in place. Furthermore, principal researchers involved with some of these projects have significant ties to the NCGIA, as in the case of the Sequoia 2000 Project.

Research under Initiative 5 has been concerned with the set of problems posed by very large spatial databases (VLSDB), particularly those anticipated to be come available in the late 1990s. Such problems are now of major interest to many institutions and researchers. For example, the EOS program of NASA, which is now underway with the awarding of contracts, will soon generate data at rates far beyond previous capabilities for processing and analysis, while the GENOME project is also characterized by significant problems relating to the storage, access and analysis of spatially indexed data.

Clearly the effective processing, storage, manipulation, and analysis of such datasets requires radically new approaches to spatial data models, spatial data structures, algorithms, and user interfaces. The many issues that are raised by such requirements have been the subject of examination by a variety of workshops, research conferences and research projects. For example, a panel that studied the achievements and opportunities for database systems indicated that the next generation of applications, such as data-intensive scientific applications, will require more sophisticated database support. Such support relates to data access and type management for large and internally complex objects, capabilities for processing large numbers of rules, and the ability to handle

new concepts such as spatial data, time and uncertainty. A second panel that examined scientific database management made the following critical observation: "*To manipulate data and produce information, a scientist needs to access data and apply analysis tools in concert. Failure to integrate the data management and analysis environments restricts the productivity of the scientist... extant systems are not integrated...due to the fact that the data management environment was created by a computer scientist and the analytic environment... by a discipline specialist.*" In relation to research projects, the Sequoia 2000 project has set itself the goal of designing and building a system that provides physical support to scientists who require access to large databases of heterogeneous elements.

In the original proposal for the NCGIA, the stated objectives of Initiative 5 were:

1. assess requirements for very large databases;
2. determine characteristic data types for remotely sensed data;
3. identify functional components for very large GIS databases and related GIS products;
4. develop methods to group components to achieve high performance;
5. build prototypes and test components.

The proposal also suggested several relatively concrete tasks on which the Santa Barbara and Orono sites could each focus.

Initiative 5 was set in motion in July, 1989 with the hosting of an open symposium on the topic of very large spatial databases, with both invited and contributed papers. A closed, specialist meeting immediately following the symposium. Oliver Gunther acted as general chairman of the symposium and Alex Buchmann acted as chair of a distinguished program committee. Support for the symposium was obtained from USGS, ORNL, and NASA. The 170 attendees comprised a diverse group of individuals from academic, commercial and government sectors. Papers were presented by four invited speakers (S.K. Chang, M.F. Goodchild, J. Nievergelt and H. Samet), and fourteen papers by individuals who had submitted papers, including work done in preparation of the initiative by Egenhofer, Frank, and Jackson; and by Frank and Barrera. Two panel discussions were also held on problems of large spatial databases and on possible solutions to these problems. The fully refereed papers were published by Springer-Verlag in their "Lecture Notes in Computer Science" series with Buchmann, Gunther, Smith and Wang as co-editors.

The success of the symposium is apparent from the fact that it has now become a biennial event: the Second Symposium on Large Spatial Databases (with Egenhofer as co-chair and co-ordinator for North and South America, and Frank, Goodchild and Smith also on the program committee) was held in August 1991 in Zurich; the Third Symposium on Large Spatial Databases will be held in Singapore in 1993 with both Frank and Smith acting as members of the program committee for this symposium; and the Fourth Symposium will be held in Orono in 1995. At the First Symposium, over 170 participants registered and 26 fully refereed papers were presented.

The specialist meeting was held immediately following the First Symposium, and was co-led by T. Smith and A. Frank. Members of the NCGIA who participated in the workshop included Barrera, Egenhofer, Ehlers and Frank from Maine and Estes, Simonett, Smith and Star from Santa Barbara. Over forty participants were present at the workshop. The attendees were an extremely experienced group of individuals from Europe and North America, representing academic,

commercial and government sectors. The workshop was structured in terms of full group sessions and several small group sessions, in which much of the work occurred. Significant contributions (and reports) were made by small groups in the following areas:

1. **Architecture of GIS databases**: Topics discussed by the participants of this group were map algebra, extensible query optimization, transaction models for GIS, integrity and validation issues for GIS, DBMS issues and GIS requirements and sample data.

2. **Concurrency control**: Topics discussed by this group included analysis and validation of long transaction models, physical concurrency on spatial indices, lock-types and consistency among multiple representations of the same data, recovery and audit trail issues.

3. **User interfaces**: Various user interface tools were discussed by this group in relation to several issues concerning queries, including zoom and pan, granularity criteria, search and browse criteria, layering criteria, launching and macro execution and transcript log and system status. The group also discussed the concept of user interface development environments.

4. **Hardware**: This group examined issues relating to user scenarios, working set characterization, exploitation of parallelism and GIS specific hardware.

5. **Object modeling**: This group examined three main areas, including object models, extensible systems and operational facilities. Research was suggested in relation to analysis of existing systems, with identification of candidates for validation; validating a set of chosen systems; and comparison of results.

6. **Acquisition, capture, integration and maintenance**: A variety of issues were examined by this group including the integration of data from a variety of sources, the construction of large databases, database maintenance techniques, source data analysis, high resolution databases and data availability.

7. **Extensibility**: The main projects identified by this group related to identification of requirements, analysis of extensible approaches, mapping of requirements onto particular systems with evaluation and comparison of evaluations to determine best approaches.

The results of the workshop were summarized by T. Smith and A. Frank in a technical report (1990) and also in an article published in the Journal of Visual Languages and Computing (1990).

In the remainder of the paper, we examine in greater detail the research agenda that resulted from the Specialist Meeting; the progress to date on the research agenda; significant augmentations of the research agenda; an assessment of the initiative; and an annotated list of publications resulting from the work. Before continuing, however, it is perhaps appropriate to note that with respect to the writing of this report, it has been very difficult to delineate those classes of activities that resulted primarily from the I5 Initiative, and those that resulted primarily from other initiatives. Different individuals are likely to have differing views as to which classes of activities are most closely related to I5. In part, this difficulty in inherent because of the close relationships between many of the initiatives, such as I2 and I5. Hence in this report there may be errors of omission or commission because of the relationships of the areas of investigation of different initiatives and the fact that the primary researchers in the I5 Initiative were also involved in closely related initiatives. In part, however, it is also due to the fact that different researchers may take differing views as

to the nature of the research issues for VLSDB. This is clear from the outcome of the Specialist Meeting, in which almost every major topic relating to VLSDB was emphasized as being of major significance for the Initiative by at least some subset of individuals. For example, it is clear that at one extreme one may focus attention on the ease with which *applications researchers* may interact with VLSDB. Hence one may focus interest on the design of database languages and interfaces for such users. At the other extreme, one may justifiably focus attention on *low-level physical support* for data storage and access. The foci of the research at both UCSB and U. Maine, although differing in emphases, have been closer to the first extreme than to the second.

2 THE RESEARCH AGENDA

2.1 The Full Research Agenda Established at the Specialist Meeting

The research agenda, as published in the NCGIA report, included the following problems:

1. Investigate how VLSB should be designed and assembled, including the question as to whether there are significant differences between large and small databases.

2. Examine how one should employ multiple sources of diverse data in order to derive metrically accurate spatial data. Are there techniques for synergistic data extraction that will allow for non-linear scaling of effort as we build larger databases.

3. Perform an analysis of source data.

4. Design and implement an improved geodetic database for scene registration.

5. Investigate database maintenance techniques for VLSDB. Particular questions of interest relate to differences between starting from scratch vs. incremental update; the possibility of storing product independent feature descriptions; the design of an intelligent gazeteer that would include a much richer set of information concerning the place name; the identification and control the loss of information during format conversion.

6. Develop, test and evaluate spatial databases for a variety of topics. The goals would be to provide researchers with a common set of databases for experimentation on a diverse set of topics.

7. Investigate the use of a high-resolution spatial database in the automated interpretation of remotely sensed data.

8. Design and implement VLSDB systems that incorporate incremental and evolutionary growth.

9. Investigate integrity and validation considerations in relation to incremental growth of VLDB. The investigation should characterize the nature and extent of data dependencies and determine appropriate concepts and methods for the maintenance of database integrity.

With respect to *user requirements and queries*

1. Develop a taxonomy of user scenarios, their subtasks, and the characteristics of these subtasks.

2. Develop a thorough characterization of the working set of data for problem solving in a GIS. This working set characterization should cover the different subtasks and characteristics of the user scenario taxonomy.

3. Characterize users in terms of the nature and classes of queries concerning the database and its contents.

4. Examine techniques involving metadata and browse that enable a user to locate data of value in a VLSDB.

5. Investigate how users think about objects and data, and how they plan their work.

6. Investigate user requirements concerning acceptable parameters of system response (including time).

With respect to *hardware and system architecture*:

1. Investigate workstation hardware impediments to the use of large GIS databases by developing or acquiring a very large and powerful workstation (500 MIPS, 500 MBytes RAM, 20 GBytes disk) and investigate its impact on GIS applications. Determine by this experimentation the effects these workstation parameters have on the GIS design, the database and working set use, and the relative importance of synchronous vs. asynchronous support for various GIS tasks.

2. Develop a new GIS workstation architecture tailored to spatial data, with two-dimensional memory access and possibly two dimensional secondary storage, and with a bus architecture to facilitate movement of two dimensional data between secondary storage and memory.

3. Investigate various database machine architectures as they might apply to processing in GIS subtasks.

4. Examine GIS subtasks for exploitation of parallelism.

With respect to *extensibility and DBMS*:

1. Examine the file system approach, particularly in relation to the design of file systems and to their efficiency.

2. Design an extensible DBMS, and in particular examine which functions should be in the physical DBMS level.

3. Investige the set of conditions that should be satisfied by an extensible GIS.

4. Investigate map algebras and extensible query optimizers based on the map algebras.

With respect to *concurrency control and transactions analysis*:

1. Analyze carefully the granularity of the data touched by GIS transactions, as well as the typical duration of transactions.

2. Analyze existing long transaction models and validate them in terms of their usefulness with respect to the requirements of geo-applications.

3. Investigate how the various access methods stand up under physical page locking and whether this will cause a drastic loss of performance (e.g. due to locking of the root page and elimination of all concurrency).

4. Investigate issues of consistency among multiple representations of the same data in a GIS.

5. Determine the best logging mechanisms for GISs (for example, before and after images or general page-level logging vs. transaction logging).

With respect to *object oriented approach*:

1. Perform a comparative investigation of OO systems, ending with a global evaluation of the best system.

2. In relation to schema, investigate the definition of basic objects and dependencies stored in the DBMS. In relation to this topic, specific aspects include aggregation hierarchies, lattices, networks, recursion, associations, is-a hierarchies, and other semantic relations and temporal relations (including data types for time).

3. In relation to views, investigate the definition of static and dynamic objects and dependencies devised from DBMS schema. Specific aspects for focused research include dynamic object definitions and operations (such as extraction, creation, deletion, and update).

4. In relation to main memory representation, investigate object buffer, operations on buffered data, structure mapping, data conversion and morphism of objects (db vs. language)

5. Investigate the basic methods of geo- and spatial applications, including definition of a (relatively small) set of domain specific methods, including geometric operators, semantic relations and consistency constraints.

6. Investigate the influence of OO approaches and the influence of operation facilities on the OO approach.

With respect to *storage and access to data*:

1. Investigate physical clustering of data on storage devices, and the value of current data structures in support.

2. Investigate the value of storing data in different formats, at different levels of resolution and at varying stages of processing.

With respect to *access to VLSDB*:

1. Investigate the nature and form of metadata to be used in VLSDB, and how it should be distributed over the architecture and used in applications.

2. Investigate the nature of the browse facilities that should be incorporated in VLSDB.

3. Investigate the design of query languages that should be employed in VLSDB.

4. Investigate benchmarks for common geo-procedures and transactions. Such transactions might include spatial form, buffering, and pan and zoom.

5. Investigate parallel implementations of geo-procedures on a variety of architectures. In particular, different programming paradigms (such as data parallelism and functional parallelism) should be investigated.

2.2 The Research Agenda As Chosen by NCGIA Researchers

Clearly the preceding list of topics covers almost the entire field of spatial database research, and as such is far too extensive in nature to form a practical research agenda. Hence the interested researchers at UCSB and U. Maine chose a research agenda by making selections from this list. The topics constituting the actual research agendas were selected on the basis of being the most appropriate for research at the two sites. The criteria for appropriateness included the importance of the topics, the interests in the topics of the researchers at the two sites. and the feasibility of making progress in the light of both strengths and resource constraints. The topics for investigation in the research agendas included:

1. Access to Spatial Data (Maine);
2. Spatial Algebra (Maine);
3. Robust Evaluation of Queries (Maine);
4. Logic-based and Object-Oriented Models and Languages for VLSDB (UCSB);
5. Metadata (UCSB);
6. Lineage (UCSB);
7. Browsing (UCSB);
8. Compression (UCSB).

It is clear that each of these topics could provide a focus for major research efforts.

3 PROGRESS ON THE RESEARCH AGENDA

We now provide a brief review of some of the progress that was achieved with respect to the research agenda. Some parts of the agenda received more attention than others partly because of the perceived importance of some items and the greater interest in such items and partly because of resource constraints that limited the matters that could be investigated.

3.1 Access to Spatial Data

During the Specialist Meeting, consensus was reached that methods for access to and selection of the spatial data are crucial for the effective use of the data collected (Smith and Frank 1990). Hence, a study of requirements for interactive spatial query languages was initiated at Orono, with a particular emphasis placed on the interaction between user and system. The results were published in an article by Egenhofer in the Journal of Visual Languages and Computing.

These requirements have provided the basis for the design of various spatial query languages. In particular, an extension of SQL was designed to handle the cartographic display of spatial query

results. It demonstrated that the incorporation of queries and display instructions into a single language would make user queries unnecessarily complex. Instead, the separation into (1) a query language and (2) a graphical presentation language allows the user to formulate less complicated queries. Articles by Egenhofer describing the language design and the architecture of such a system have been published in Cartography and Geographic Information Systems and IEEE Knowledge and Data Engineering, respectively.

Based on the experience with extending SQL for spatial data handling, studies were conducted concerning the usefulness of SQL as a foundation for an interactive spatial query language and a number of serious deficiencies were identified. Particular difficulties include the incorporation of the necessary spatial concepts into SQL and the limitation of the relational framework when dealing with geographic data such as complex objects, object identity, meta queries, and non-first-order logic. Egenhofer presented different aspects of this assessment at various computer science conferences and the NATO ASI in Las Navas and a comprehensive version has been published in the International Journal of Geographical Information Systems.

A. Frank investigated new spatial query languages, particularly suitable for the exploratory access geo-scientists need, published at a European workshop on Database Management Systems for Geographical Applications. He proposed an iconic language using "data cubes" (for data sets) and "maps" (for the graphical presentation of data sets) so that "data cubes" can be explored by dragging them over "maps". This separation is in concert with the results of the design of Spatial SQL where separate languages were used to formulate the queries and display instructions.

3.2 Spatial Algebra

During the specialist meeting in Santa Barbara, a need for a better understanding of the geometric operations used in spatial databases became evident. In cooperation with the University of Bremen (Germany), the Maine site used an algebraic formalization of a "map algebra". Originally, such a concept had been described informally by Dana Tomlin (Ohio State University) as a set of operations, which can be applied to a "map layer" producing another layer or to a set of layers to combine them into a new one. These ideas were formalized using a single method (algebraic specification) describing the types and their arguments and giving axioms that specify their behavior. This formal description of the map algebra permitted the investigation of viable strategies for optimizing the combination of several overlay operations. Preliminary results were published at Autocarto 10. This work has fostered further investigations of properties of different spaces by using algebraic methods, in relation to Initiative 10.

3.3 Robust Evaluation of Queries

Geographic databases may keep versions of the same map with different levels of precision. Hence evaluation of queries with requests for the aggregation of many detailed values in a database are of particular importance for GIS. They occur whenever a sum or a count for an area is requested and the individual data elements are stored. The more aggregated and less precise a representation is, the fewer instances are recorded and the less storage is occupied. For certain users and tasks, a result with less precision may be usable. For example, the display of an overview map may be sufficiently detailed based on the selection of a few significant objects. A trade-off between precision of the answer and reponse time asks for optimization. If one includes the factor of time available to perform a certain operation, it is possible to treat this as a trade-off between precision of the result

and processing time: the more time available, the more precise one can determine the result. The requirements for such a system are (1) to perform incremental evaluations and (2) to assess how much a partial result deviates from the final, most precise result so that users can be information about the limitations of the answer.

Barrera and others have developed algorithms for determining the incremental solution of relational algebra operations and for estimating the deviance. A paper by Barrera, Egenhoffer and Frank describing these results was presented at the Symposium on Spatial Data Handling held in South Carolina, 1992.

3.4 Logic-based and Object-Oriented Models and Languages for VLSDB

The development of high-level languages that support both the database activities and the more general modeling activities of scientific researchers may be viewed as both a specific topic of research and as a thread that links many of the issues in VLSDB together. There are many important research issues involving the syntax, semantics, expressiveness, usability and efficient support of such languages that have close relations to most of the issues in the I5 research agenda, including access, spatial algebra, robust evaluation, metadata, browsing and lineage. Hence the development of such languages has been the central focus of the UCSB effort in the area of VLSDB.

T. Smith, together with R. Ramakrishnan (University of Wisconsin) and A. Voisard (INRIA) worked on data models and languages for defining and modeling complex spatial objects and their behavior. These models and languages are based on logic and may be viewed as extensions of relational models and languages, as well as containing object-based aspects. Some of the more general ideas underlying this research were presented by T. Smith in an article in *Geographical Information Systems: Principles and Applications* (1991).

The initial research led directly to a three year award for over $600,000 from NSF in the area of scientific databases to T. Smith and six other investigators, including Drs. A. El Abbadi, D. Agrawal and J. Su, of the CS Department at UCSB; J. Dozier, of the Department of Geography at UCSB (who is also a PI on the Sequoia 2000 project); R. Ramakrishnan, of the CS Department of the University of Wisonsin; and T. Dunne, of the Department of Geology at University of Washington. There is also three years of support from NASA for a graduate student who is working on the project. The proposal was in large part based on work presented in a paper by Smith, Ramakrishnan and Voisard on object-based data models and deductive languages presented at a workshop held in Capri in 1991. The initial part of the research carried out under this award must be considered as an integral part of the I5 research initiative research that was performed at Santa Barbara. At this point, we describe those aspects of the research that were performed before the formal closing of I5. Below, we present additional information concerning post-I5 developments.

The long-term goal of this research is the design, analysis and evaluation of computational systems that both unify the modeling and database operations of many areas of scientific research and lead to greater research efficiency for users. Particular interest relates to scientific applications, such as *large-scale environmental modeling*, that are intensive in terms of both computation and data, and that involve the definition and manipulation of complex, spatio-temporal entities. The data models and languages are being developed in relation to computational environments in which data access and tools for analysis are integrated in order to support data-intensive and numerically intensive modeling activities in a class of scientific database systems. They involve a simple model of scientific activity in order to motivate the development of a new conceptual data model for complex,

spatio-temporal objects and a deductive, object-oriented language to support the definition and manipulation of such objects.

A focus of the investigation is the design of *modeling and database systems* (MDBS) that provide explicit support for complex spatio-temporal entities, properties and relationships, and transparent support for data access from large, heterogeneous and distributed data sources. An important goal of the research is the design of a family of formal languages that express scientific modeling and database activities at a level allowing scientists to focus on scientific rather than on computational issues. It is believed that such languages, termed *modeling and database languages* (MDBLs), should permit scientists to define and manipulate entities at all levels of *abstraction* from simple datasets to large, mathematical models. In particular, such languages should support the definition and manipulation of *effective representations* of domains of complex, spatio-temporal entities, properties and relationships. They should also possess efficient computational implementations and translate easily into a variety of interface languages.

The language design is currently being based on the requirements of a group of EOS investigators whose domain of study is the hydrology, limnology and geomorphology of the Amazon drainage basin. The overall aim of these investigators is to understand the effects of land-use change, the effects of climatic and surface features, and the interrelationships between the hydrological and biogeochemical cycles in the Amazon Basin. The Amazon project is representative of many EOS projects in the sense that the phenomena under investigation are complex; the dataset and processing requirements are large; while there are non-trivial requirements for scientific interaction. The mode of investigation in this domain is changing from one of simple, lumped-system modeling to one based on distributed system modeling. Such an approach requires the definition and manipulation of many domains of complex spatio-temporal entities.

Some preliminary results of this research were presented in a paper at the Symposium on Spatial Data Handling held in South Carolina (August, 1992). The main contribution was the justification, definition and application of a *term definition language* (TDL), which is a metalanguage in which users may define domains of complex, spatio-temporal entities and point to associated procedures. Such domains and procedures together form an important component of an MDBL. There are important and non-coincidental correspondences between TDL/MDBL and current research in databases and database languages, as, for example, between TDL and the structural data model of O^2, as well as the functional database languages of Buneman; and between MDBL and current, logic-based, object languages, such as COL or STARBURST SQL.

3.5 Metadata

The issue of metadata is very important in relation to accessing the appropriate information in a VLSDB. This topic was studied by T. Smith and N. Trivedi in the latter's MS thesis. Details of the research are provided in an NCGIA Technical Report. A conceptual framework for integrated metadata management in large spatial databases was constructed. The primary function of this framework is to allow for the definition, location and control of metalevel information about the underlying database. The framework provides for a set of core metadata components and allows for addition of any auxiliary metadata that the user might want to define. The framework would support feature based retrieval as well as interactive browsing of metadata. The emphasis is on flexibility, extensibility and ease-of-use. The goal is integrated management of all kinds of metadata. The report gives an overview of semantic modeling of spatial data followed by a conceptual model for

metadata. The basic idea of the conceptual model is to classify the database entities of interest into data, process and environment entities. Corresponding to this, the metadata consists of metadata, metaprocess and metaenvironment entities. The report proposes a *forms* mechanism to manage metadata. A set of basic operations for manipulating forms and catalogs is described. A case-study of metadata in a conventional GIS environment is presented. This is supported by the preliminary version of the schema for the Condor Database Project at the University of California at Santa Barbara.

3.6 Lineage

Lineage, which may be viewed as a special case of metadata, involves keeping track of the origins of intermediate data layers. Lineage is an important issue for VLSDB, as for example, in the case of assessing error propagation, and has been studied by D. Lanter and students at UCSB. Lineage documentation specifies an application's source data, transformations, and input/output specifications. Such information is inherently causal, communicating the theory embodied in a GIS application and the meaning of its product. A number of techniques for automating lineage information were examined by Lanter, but none were found to be capable of documenting data lineage. Lanter has presented a conceptual design of a meta-database system for documenting data sources and GIS transformations applied to derive cartographic products. Artificial intelligence techniques of semantic networks are used to organize input-output relationships between map layers and frames to organize lineage attributes characterizing source, intermediate, and product layers. An illustrative example indicated that a lineage meta-database enables GIS users to engage in source assessment throughout their analysis of spatial data sets.

3.7 Browsing

Browsing capabilities are closely related to the usual concepts of metadata as tools for accessing appropriate information in VLSDB. This topic has been studied by J. Star and students at UCSB. The research has asked questions about user requirements, and looked at the determination of a minimum set of processing algorithms to extract very small browse images from large digital images.

3.8 Compression

The volumes of data in a VLSDB can become so large that lossy compression must be considered. The problem was studied at UCSB by a team involving A. Gersho and S. Gupta, J. Star and T. Smith and students. Most of this work was based on the vector quantization techniques that Gersho has been developing for a number of years in relation to image storage. In particular, Gupta and Gersho (1990) examined vector quantization compression techniques with respect to multispectral satellite images of the earth, which consist of sets of images obtained by sensing electromagnetic radiation in different spectral bands for each geographical region. They proposed a new compression method for such data sets, in which a small subset of image bands is initially vector quantized. The remaining bands for the same spatial region are estimated from the quantized images by a nonlinear predictor which is optimal for the mean squared error distortion measure. The residual (error) images are conditionally encoded at a second stage based on the magnitude of the errors. This scheme exploits both spatial (by vector quantization) and spectral (by nonlinear prediction)

correlation inherent in multispectral images. Simulation results on an image set from the Thematic Mapper with 7 spectral bands were presented and a substantial improvement was obtained by using the nonlinear predictor over the optimal affine predictor. Image compression ratios between 20 to 30 are achieved with remarkably good image quality.

4 SIGNIFICANT AUGMENTATIONS OF THE RESEARCH AGENDA

The research agenda, as adopted by NCGIA researchers and listed above, relates to a set of interconnected but discrete topics. It has become clear that many of the issues relating to large spatial databases are in fact highly interconnected, and that studying each in isolation is often a rather artificial exercise. This fact is reflected in two activities that have arisen *as a direct result of the I5 Initiative*, namely the NSF-funded scientific database research at UCSB and the ESPRIT-funded research at Vienna in relation to the characterization of spatial data. These activities have, in a sense, continued the lines of research began in I5 after the formal completion of the I5 Initiative. Since they may be viewed as involving significant augmentations of the research agenda of the I5 Initiative, we believe that it is of value to summarize each of the projects.

4.1 The NSF-funded Project at UCSB

An overview of the project and initial results was presented at the AAAS meetings in Boston (February, 1993) and was also published in the Bulletin of Data Engineering (Smith, Su, Agrawal and El Abbadi, 1993). As noted above, the overall goal of the scientific database research project at UCSB is to design and develop computational support that will permit scientific investigators to achieve their scientific goals more efficiently (earth science research teams may focus up to 50% of their attention on computational issues that are irrelevant to their scientific research). Since it is critical, however, to provide support that will be adopted and used by scientists, a first step of the research has been to understand the nature of the earth science investigations and the computational issues involved in such investigations. A second step involves designing and investigating *modeling and database systems* (MDBS) that provide explicit, high-level support for: (1) iterative model development; (2) database construction, maintenance and access; (3) multi-investigator research projects.

The general approach to achieving these goals involves: (1) close, collaborative research with the scientists of a specific and representative project, in this case the Amazon project; (2) the design, investigation and prototyping of an MDBS whose functionality is based on an *appropriate model of scientific activity in which modeling and database activities are closely linked*; (3) a top-down design for an MDBS that is based on a specification for a high-level *modeling and database language* (MDBL). Such a language should be capable of expressing, in simple and natural terms, the greater proportion of the scientists' requirements with respect to modeling and database activities.

A key element of the research involves basing the design of an MDBS on an appropriate model of the goals and activities of scientists. The investigators have therefore developed a model scientific investigations as activities in which scientists (1) construct, evaluate and employ a large set of R-domains for representing conceptual entities; (2) construct, evaluate and employ representations of many transformations between these R-domains; (3) construct instances of R-domain elements; (4) apply sequences of specific transformations to specific sets of R-domain elements. An MDBS is

being designed that adheres to this view of scientific activity and that, in particular, supports these four general classes of activity. Central to the MDBS is a high-level, largely declarative modeling and database language (MDBL). In relation to the iterative development of scientific models of phenomena and the associated activities of database creation and access, MDBL may be employed by investigators to construct and apply an extensible and potentially *very* large collection of R-domains of elements and associated transformations. Such a "lattice" of domains may be viewed as representing a relatively "complete" set of concepts that a scientist needs for modeling some set of phenomena, from "low-level" datasets to "high-level" models. MDBL supports the declarative specification and efficient construction of transformations associated with the domains; the creation of domain instances; and the application of specific domain transformations to instances, or sets of instances, of domain elements. MDBL also supports the representation of database activities in which scientists may view any dataset as an element of some domain, rather than as a collection of files that may be distributed in the current computational environment, and access datasets by content. Since MDBL is designed to permit scientists to express most of their computational activities in a simple and natural syntax, the language is not computationally complete and access to other more complete languages is therefore facilitated.

The database support underlying MDBL incorporates the following general features: (1) uniform structuring and organization of the data that is independent of the physical implementation of the system and the distributed nature of the database (in particular, users see data in terms of named "virtual" datasets rather than in terms of files); (2) simple access to datasets that avoids problems related to I/O (for example, access to portions of datasets may be made in terms of domain names and declaratively expressed constraints on the values of the elements sought); (3) construction of, and access to, transformations on the domains; (4) support for concepts such as *project* and *database view*, including support for concurrent access to datasets within a project, that are independent of the system implementation; and (5) automatic updating of the database. Another degree of complexity arises as a result of the physical distribution of the domains. Techniques are being developed to provide a logically centralized view of a domain that may be distributed spatially. In particular, as a first step towards prototyping MDBS, the investigators are using existing file-systems as an underlying store for the domains in MDBS as well as the Prospero file-system to provide an integrated view of an R-domain that may have been represented in terms of multiple files distributed over a network. Although file-systems impose a maximum size for a single file, the size of a domain is not restricted since it is represented in terms of a (theoretically) unlimited number of files. Furthermore, the Prospero system allows files corresponding to a domain to be organized in a variety of ways that include temporal, spatial, spatio-temporal and content-based indexing.

The sharing of information among scientists involved in various projects appears to be important and is being investigated as part of the database support underlying MDBL, and involves issues relating to concurrency. Traditional database systems control concurrent executions by requiring that such interactions to the databases be atomic. However, serializability is very restrictive in scientific databases since interactions in the database may have very long durations. Since maintaining atomicity of long interactions imposes long delays on users, the investigators have developed a model based on the notion of "relative atomicity" of interactions. In this model, the users explicitly specify the ways in which their interactions to the database should be interleaved with the interactions of others. For example, there may be no atomicity restrictions on interactions that are initiated within a project whereas these interactions must be atomic with respect to interactions

of other projects.

The prototyping and experimental components of the research are in part intended to capture the nature of computation in scientific research and in part to investigate and test the conceptual MDBS that are being designed in order to support such research. A first component of the experimental research involves the development of MDBL and its appropriateness for scientific database and modeling applications. In order to capture the computational "requirements" of the scientific investigators on the Amazon project, complete sequences of activities relating to model development and database access have been captured in MDBL, and in other more complete languages where necessary. This activity has involved the iterative design and representation in MDBL of an appropriate lattice of abstract and concrete R-domains as well as expressing, in MDBL, several examples of specific scientific investigations. For situations in which a more complete language than MDBL is required, use is being made of the deductive database language CORAL, which combines the general logic programming paradigm and database manipulation, and other programming languages. A large number of domains and associated transformations have been designed and implemented and several application examples that were previously written have been rewritten in MDBL and CORAL. A second component of the experimental research relates to providing a network transparent file system layer to support this database design by using the PROSPERO system to provide a *virtual* file system over the network.

4.2 The Esprit-funded Research at Vienna

The research funded by ESRIT and being carried out by researchers at the University of Vienna in collaboration with other partners includes:

1. the analysis of user requirements, in terms of the use of spatial concepts in spatial data management systems;

2. the analysis of typical parameters characterizing spatial data sets in typical applications and the analysis of user application requirements in real case studies.

The project consists of three interrelated subprojects that relate to:

1. user requirements and case studies;

2. data models and languages;

3. systems architecture and prototyping.

We now briefly summarize each of these subprojects.

The investigation of user requirements and case studies involves two major steps, namely the definition of what must be measured (i.e. specific values and parameters on spatial data) and case studies (i.e. actual measures on real life data files and actual experimentation on real life application problems). The group at University of Vienna will define accurately what is to be measured and describe procedures and portable programs to measure these characteristics in data files. Possibly, this will lead to the definition of benchmarks, performance criteria and parameters for evaluating various GIS software. Research partners who handle real geographic data will in the second step adapt these programs to their particular data structures and measure actual values for typical data files of their applications. Research partners interacting with end-users will cooperate with them

in the definition of how technological solutions can be used to solve real life problems. An initial classification of the parameters to be estimated through measurement involves four categories, namely data volume; the distribution, size and topological characteristics of spatial objects (in relation to topology and distribution); the modeling of geometry; and changes in data over time.

The investigation of data models and languages involves the objectives of defining extensions to the relational model in order to deal with geographical information; modeling network structure, to define methods for representation of heterogeneous collection of objects; analyzing modeling problems regarding interaction between mathematical models and spatial DBMS; and investigating topological data models. While much effort stills needs to be expended in relation to general models for geographic data and in relation to understanding how to represent geometric information, several new modeling issues are being studied, including models for geometric precision and problems of lineage of geometric information; the modeling of (spatial) networks and heterogeneous collections of objects; the interfacing of mathematical models to geographic databases. The following tasks have been defined as a basis for studying these problems:

1. the formalization of simplicial complexes and chains for the modeling of 2D geometry;
2. the development of models for lineage and time management;
3. the investigation of graph databases and heterogeneous collections of spatial objects the interfacing of mathematical models to geographical databases.

In relation to the issue of system architectures and prototyping, the main research objectives are to study: methodologies for embedding data structures for representing graphs and for heterogeneous collections of spatial objects into database system architecture; data structures for partitions of the plane and persistent storage management systems; cooperation between an extensible DBMS and an external process providing geometric computation as a service; the use of the O2 system as the basis for experiments with real life data; extensions to offer capabilities for interfacing mathematical models for environmental analysis and planning with a given relational system with spatial data types.

5 ASSESSMENT

This section presents an assessment of the initiative. Part of the assessment is made in general terms, while part is organized according to the five criteria for initiative assessments established by the NCGIA Board of Directors. It is to be emphasized, however, that much of the written output that resulted from the VLSDB Initiative was quite preliminary, and one should interpret these results accordingly.

Relative to the scale and scope of the research activities of such large scale projects as EOS, GENOME and Sequoia 2000, it is now clear that Initiative 5 was overly ambitious in terms of its scope and could, at most, hope to have only a modest impact on major research projects that involve VLSDB and that exist largely outside of the usual Geographic and GIS communities. Nevertheless, we believe that the initiative did have a major impact on the database community. Since the start of the initiative, the topic of "geographic databases" has become an acknowledged research issue among computer science researchers. The specialist meeting, in particular, fostered the idea of real cooperation between GIS and computer science researchers. Participants at that meeting identified

the need for GIS researchers to provide detailed specifications of GIS database requirements so that computer scientists could provide corresponding technical solutions. Similar support for cooperation was initiated from other computer science areas, such as temporal databases and human computer interaction, at later NCGIA-sponsored meetings (Temporal Database Workshop at Orono, Interfaces to Geometry Workshop at SIGCHI). Since then, NCGIA researchers have produced a number of database requirements articles, and have increased their cooperation with database researchers. A. Frank and T. Smith were both involved in the European Basic GOODS workshops, an activity in which database researchers focussed on geographic object-oriented database systems. More and more prestigious database journals and conferences solicit research results on spatial and geographic databases. For example, a proposal by Frank and Egenhofer for a tutorial on Geographic Databases was selected for the 1992 Conference on "Extending Database Technology". Over the next years, we are expecting further involvement of, cooperation with, and support from these communities.

Discussions among participants at the Initiative 5 Workshop provided the groundwork for a Special Interest Group on Spatial Information Systems (SIG Spatial) within the Association of Computing Machinery (ACM). Such a group would provide the focus for all those disciplines building on technical advancements for spatial information systems. SIG Spatial is being planned as a group working closely with AAG's SIG GIS and bridging between geography and the different computer science disciplines which can contribute to overcoming GIS impediments.

We now consider each of the questions posed by the Board of Directors.

5.1 RESEARCH ACCOMPLISHED: What do we know now that we did not know before the questions addressed by the initiative?

We may view our new knowledge in terms of relatively concrete answers to specific questions and also in terms of more general knowledge. In relation to concrete answers to specific questions raised concerning the seven topics that formed our research agenda, we now have *preliminary* knowledge that indicates how we might, in relation to:

1. *spatial algebra*, specify in algebraic terms, as in other areas of computer science, map layers and the operations by which they are manipulated.

2. *robust evaluation of queries*, use query execution strategies that trade off execution time and precision of the results;

3. *logic-based languages*, employ a *useable* logic-based language that involves modeling phenomena over a large range of domains of spatio-temporal elements as well as the construction of domains of such elements and operations on such domains using a schema definition language; the requirements of scientists with respect to such language; and issues involved in supporting and implementing such a language;

4. *metadata*, construct a framework for metadata management in large spatial databases (for example, this has already been accomplished in the MDBS project);

5. *browsing*, employ one set of user requirements concerning metadata and a minimum set of algorithms for extracting browse data from a set of images;

6. *lineage*, design and implement semantic networks that keep track of lineage data in large spatial databases;

7. *compression*, employ a new vector quantization technique for compressing image and digital cartographic data.

Hence, we have for each topic in our research agenda some applicable and concrete knowledge that was not available before the beginning of the initiative.

We also have acquired some very general ideas that relate as much to other initiatives as they do to I5 and that could have important implications for the manner in which we conduct future research in these areas. For example,

1. the value of logic-based and related object-based languages together with the importance of constructing systems that are based on a strong theoretical framework and that combine database and modeling languages in a natural way became apparent during the I5 research at UCSB. *In particular, this idea provides a framework for integrating modeling activities into GIS and VLSDB systems in a manner that is far more systematic and far easier than the current ad hoc approaches permit.*

2. The concept of a lattice consisting of a large number of spatial and non-spatial domains of elements with associated sets of transformations, in the context of the structure that we have imposed on such a lattice, provides a major unifying basis for designing and tailoring systems that are applicable to a large number of domains of application.

These ideas do not appear to have been explored in the context of GIS.

5.2 RESEARCH AGENDA DEVELOPMENT: How has the research agenda been affected by this initiative?

The performance of the research has had a significant impact on the research agenda. Clearly, as research progressed on each of the seven topics, the issues to be explored have generally become clearer, but also broader in some cases and narrower in others. Of more significance, however, is the fact that several issues that were brought up as problems at the specialist meeting, but which were not incorporated into the research agenda, have found their way into the current research agenda. The NSF funded database project at UCSB, for example, has been forced to consider, at least to some degree, the following issues that were raised at the specialist meeting:

1. Design and implement VLSDB systems that incorporate incremental and evolutionary growth.

2. Investigate integrity and validation considerations in relation to incremental growth of VLDB. The investigation should characterize the nature and extent of data dependencies and determine appropriate concepts and methods for the maintenance of database integrity.

3. Develop a taxonomy of user scenarios, their subtasks, and the characteristics of these subtasks.

4. Develop a thorough characterization of the working set of data for problem solving in a GIS. This working set characterization should cover the different subtasks and characteristics of the user scenario taxonomy.

5. Characterize users in terms of the nature and classes of queries concerning the database and its contents.

6. Examine techniques involving metadata and browse that enable a user to locate data of value in a VLSDB.
7. Investigate how users think about objects and data, and how they plan their work.
8. Design an extensible DBMS, and in particular examine which functions should be in the physical DBMS level.
9. Investigate map algebras and extensible query optimizers based on the map algebras.
10. Analyse carefully the granularity of the data touched by GIS transactions, as well as the typical duration of transactions.
11. Analyze existing long transaction models and validate them in terms of their usefulness with respect to the requirements of geo-applications.
12. Investigate how the various access methods stand up under physical page locking and whether this will cause a drastic loss of performance (e.g. due to locking of the root page and elimination of all concurrency).
13. Investigate issues of consistency among multiple representations of the same data in a GIS.
14. Determine the best logging mechanisms for GISs (for example, before and after images or general page-level logging vs. transaction logging).
15. In relation to schema, investigate the definition of basic objects and dependencies stored in the DBMS. In relation to this topic, specific aspects include aggregation hierarchies, lattices, networks, recursion, associations, is-a hierarchies, and other semantic relations and temporal relations (including data types for time).
16. In relation to views, investigate the definition of static and dynamic objects and dependencies devised from DBMS schema. Specific aspects for focused research include dynamic object definitions and operations (such as extraction, creation, deletion, and update).
17. In relation to main memory representation, investigate object buffer, operations on buffered data, structure mapping, data conversion and morphism of objects (db vs. language)
18. Investigate the basic methods of geo- and spatial applications, including definition of a (relatively small) set of domain specific methods, including geometric operators, Semantic relations and consistency constraints.
19. Investigate physical clustering of data on storage devices, and the value of current data structures in support.
20. Investigate the value of storing data in different formats, at different levels of resolution and at varying stages of processing.

While each of these topics will obviously not be considered in great depth, the UCSB researchers are being forced to give some consideration to each topic as required by the task of designing and building a prototype system. Finally, it is interesting to note that of the original five objectives that were laid out in the 1987 proposal for the NCGIA (see above) three have turned out to be of major significance, namely:

1. assess requirements for very large databases;
2. identify functional components for very large GIS databases and related GIS products;
3. build prototypes and test components.

5.3 CONTRIBUTION TO GIS EDUCATION: How has the education of the GIS scientist been enhanced by the initiative?

While it is very difficult to estimate the contribution that the initiative has had on GIS education, it is probably fair to say that it has focussed the attention of more members of the GIS research community than would have otherwise been the case, on issues that have, in the past, incorrectly been considered as the province of Computer Science. It is clear that the seven topics of the research agenda that were chosen for investigation would have been thought of as "pure" computer science issues before the initiative. It is clear from the results of the initiative research, however, that these items raise very important issues for a large number of GIS researchers. Each of the research issues has introduced important ideas and terminology that were not generally current in the GIS field. As a result of public and private presentations of the research materials of the initiative, such ideas now have much greater currency. For example, it is now quite possible that the concept of high-level languages such as MDBL and its associated schema and procedure definition languages, may have an important role to play in the introduction of significant modeling capabilities into GIS and VLSDBS in general.

5.4 SCIENCE POLICY: What recommendations would the NCGIA make in this area?

Several important issues with respect to science policy are raised by the research of Initiative 5. These issues include:

1. the issue of standards for GIS and spatial database systems, which is only now beginning to be addressed in any significant manner with the adoption of the SDTS. It is reasonable to assume that standards for metadata and lineage information should be examined. In particular, since metadata is essentially data about data that has implications for users as well as for systems, a standard for content and format for such data could well prove important. The USGS is currently involved in an investigation of such a standard. Since lineage information is really a special case of metadata, the same comments apply.
2. Compression in the context of GIS is an issue that has barely been addressed by researchers outside of those involved in the Initiative 5 research. With the large volumes of digital map and image data that are becoming available, compression becomes an important issue. The main question of a policy nature concerns the issue of lossy compression and the suitability of various forms of compressed spatial data for various scientific and policy questions. This is an issue that was not addressed in the initiative research and deserves significant attention. Before recommendations can made about the appropriateness/inappropriateness of different forms of compression in different contexts, it is clear that we must understand this issue much better. For example, while we now know that many forms of lossy compression do not have a serious impact on human perception, we have little idea of how such techniques affect the conclusions that may be drawn in various scientific analyses of spatial data.

5.5 THE RESEARCH INITIATIVE PROCESS: What were the strengths and weaknesses of the research initiative process in facilitating the research in the initiative?

The flexibility of the research initiative process proved to be a source of strength. This flexibility, for example, permitted an initial symposium *and* the specialist meeting to be held in a back to back manner at the beginning of the initiative. This greatly increased the efficiency of the process of putting together a research agenda, and served as a model for other initiatives. Another strength arising from the flexibility of the initiative process was the manner in which it permitted research issues to be addressed in a modular manner by subgroups of researchers who could work on their particular research agenda items in a relatively independent way. The flexibility also allowed researchers to follow interesting directions that could not be foreseen at the beginning of the process. Such flexibility permitted, for example, researchers at UCSB to focus on the issues of high-level database and modeling languages for supporting scientific research. This focus led to a successful, multi-year proposal to NSF for the funding of a research team that was composed of researchers from a number of universities, including two outside of the NCGIA site universities.

On the other hand, the flexibility was also a weakness. For example, although there was a mechanism in place to formulate a research agenda in the specialist meeting, there was no real mechanism for selecting a *realistic* subagenda that would provide the basis for NCGIA research after the meeting. As a result, so many large and important topics were raised at the meeting that NCGIA researchers could not have any hope of addressing more than a small portion number of them. As a consequence, the selection of the subagenda for NCGIA research was somewhat ad hoc. In fact, the Santa Barbara and Orono researchers independently chose their own research subagendas, and proceeded to work in relatively independent modes. While there is no indication that this led to significant problems relating to the quantity or quality of the research, one may suspect that a more disciplined approach to the actual research subagendas for NCGIA researchers might have proven beneficial.

6 ANNOTATED LIST OF PUBLICATIONS RESULTING FROM THE RESEARCH

We note that this section includes publications that some may view as more appropriate to other initiatives while it excludes publications that some may view as appropriate. Unfortunately, there is no simple solution to the issue of what to include and what to exclude.

6.1 Articles in Refereed Journals

Egenhofer, M., (1991). Extending SQL for graphical display. Cartography and Geographic Information Systems, v 18(4), pp. 230- 245.

A language has been designed to describe the cartographic display of query results in a geographic information system. Its syntax is based on SQL, the standard query languages for relational databases. The novel approach is the syntactical separation of database query and display specifications into a query language and graphical representation language, respectively. Spatial SQL introduces spatial data types and the corresponding spatial relationships, allowing users to inquire about spatial objects

in the familiar SELECT–FROM–WHERE form, extended by spatial conditions. The cartographic display of spatial objects selected is directed with the Spatial SQL-based graphical presentation language, so that complex graphic descriptions can be formulated in a language very similar to SQL. The graphical presentation language contains commands to direct the display of objects, spatial context, the query window, map scale, etc. This allows users to formulate separately queries and display specifications which are integrated during query processing so that an optimized execution strategy in a single step can be achieved. It overcomes the inherent problem of previous spatial query languages which concentrated on the retrieval of data from the database and either tried to integrate the cartographic display into the actual user query or used only default renderings.

Egenhofer, M., (in press). Spatial SQL: A Query and Presentation Language, IEEE Transactions on Knowledge and Data Engineering.

Recently, attention has been focused on spatial databases which combine conventional and spatially related data such as Geographic Information Systems, CAD/CAM, or VLSI. A language has been developed to query such spatial databases. It recognizes the significantly different requirements of spatial data handling and overcomes the inherent problems of the application of conventional database query languages. The spatial query language has been designed as a minimal extension to the interrogative part of SQL and distinguishes from previously designed SQL extensions by (1) the preservation of SQL concepts, (2) the high-level treatment of spatial objects, and (3) the incorporation of spatial operations and relationships. It consists of two components, a query language *to describe what information to retrieve and a* presentation language *to specify how to display query results. Users can ask standard SQL queries to retrieve non-spatial data based on non-spatial constraints, use Spatial SQL commands to inquire about situations involving spatial data, and give instructions in the Graphical Presentation Language GPL to manipulate or examine the graphical presentation.*

Egenhofer, M. and A. Frank, (in press). Object-Oriented Modeling for GIS, Journal of the Urban and Regional Information Systems Association.

The data model upon which most of today's commercial database management systems are based has shown to be insufficient for Geographic Information Systems (GISs). The recently promoted object-oriented model *provides some useful tools for data abstraction and data structuring, which augment the conventional tools and overcome some deficiencies inherent to the traditional relational model. In particular, the concepts of* complex objects *and* pertinent operations *are more powerful modeling methods than the currently popular structure of relational tables and relational algebra. This survey article presents the concepts of object-oriented modeling applied to geographic data and demonstrates their impact on future GISs.*

Egenhofer, M., (1992). Why not SQL! International Journal for Geographical Information Systems, Vol. 6, No. 2, pp. 71-85.

The application of traditional database query languages, primarily SQL, for Geographic Information Systems (GISs) and other non-standard database applications has been unsuccessfully tried; therefore, several extensions to the relational database query language SQL have been proposed to

serve as a spatial query language. *Here it is argued that the SQL framework is inappropriate for an interactive GIS query language and an extended SQL is at best a short-term solution. Any spatial SQL dialect has a number of serious deficiencies, particularly the patches to incorporate necessary spatial concepts into SQL.*

Egenhofer, M. and A. Frank, (1990). LOBSTER: Combining AI and Database Techniques, Photogrammetric Engineering & Remote Sensing, Vol. 56, No. 6, pp. 919-926.

The powerful logic-based concept of Prolog has been integrated with a database suitable for spatial data handling to form a database query language that is more flexible and powerful than the currently used SQL. This experimental implementation, called LOBSTER, allowed researchers to explore a number of areas of a GIS. Examples from object-oriented modeling, geomorphology, and query optimization show the application of such a language. Problems encountered during the application of LOBSTER include the absence of consistency checking during input of rules and facts, and the lack of appropriate techniques to detect cyclic rule definitions. Nevertheless, the experimental implementation showed that these techniques were extremely valuable for GIS.

Gao, P. and Smith, T.R. (1989). Space efficient hierarchical structures: relatively addressed compact quadtrees for GIS. Journal of Image and Vision Computing 7(3): 173-177.

Traditional pointer-based quadtree data structures are generally viewed as inferior to linear quadtrees when used in GIS. This paper presents an improved pointer-based quadtree called the relatively addressed compact quadtree. Storage requirements are comparable with those for linear quadtrees. A memory management scheme is also proposed for managing the pointer-based quadtree in secondary storage.

Lanter, D.P., (1991). Design of a Lineage-Based Meta-Data Base for GIS, Cartography and Geographic Information Systems, Vol.18, No.4, pp.255-261.

A conceptual design is presented for a lineage meta-data base system that documents data sources and geographic information system (GIS) transformations applied to derive cartographic products. Artificial intelligence techniques of semantic networks are used to organize input-output relationships between map layers, and frames are used for storing lineage attributes characterizing source, intermediate, and product layers. An example indicates that a lineage meta-data base enables GIS users to determine the fitness for use of spatial data sets.

Lanter, D.P., (1993). A Lineage Meta-Database Approach Towards Spatial Analytic Database Optimization. Cartography and Geographic Information Systems.

This work demonstrates how increasing levesl of intelligence can be added to commercial GIS. The lineage knowledge representation introduced in an earlier article (Lanter 1991) provides a basis for encoding the logic applied within specific spatial analytic applications. This knowledge is general, reusable and extendible to solving many significant problems in geographic information processing. In the present study the lineage knowledge representation is applied to automatically: differentiate between source and derived layers, pick an optimal spatial analytic database configuration, and gen-

erate applications tailored to the contents of the spatial analytic database.

Menon, S. and Smith, T.R. (1989). A declarative spatial query processor for GIS. Photogrammetric Engineering and Remote Sensing 55(11): 1593-1600.

The design and implementation of a declarative GIS query processor capable of extracting the locations of complex objects from a spatial database is described. The processor extracts objects from the database in a single, automated step. Search is based on an efficient query processor, named forward constraint propagation, that integrates spatial constraint propagation, geometric search using hierarchical data structures, and an effective heuristic used in solving constraint satisfaction problems.

Smith, T.R. and K. Park (1992). Algebraic Approach to Spatial Reasoning. International Journal of Geographical Information Systems 6(3): 177-192.

A simple, exemplary system is described that performs reasoning about the spatial relationships between members of a set of spatial objects. The main problem of interest is to make sound and complete inferences about the set of all spatial relationships that hold between the objects, given prior information about a subset of the relationships. The spatial inferences are formalized within the framework of relation algebra and procedurally implemented in terms of constraint satisfaction procedures. Although the approach is very general, the particular example employs a new "complete" set of topological relationships that have been recently described in the literature. In particular, a relation algebra for these topological relations is developed, and a computational implementation of this algebra is described. Systems with such reasoning capabilities appear to have many applications in geographical analysis, and could be usefully incorporated into GIS and related systems.

Smith, T.R. and Frank, A.U. (1990) Report on Workshop on Very Large Spatial Databases. Journal of Visual Languages and Computing 1(3): 291-309.

On July 19-22, the National Center for Geographic Information and Analysis held the specialist Meeting of the Research Initiative on Very Large Spatial Databases (VLSDB) at Santa Barbara, CA. At this workshop, 42 participants from the U.S. and Europe discussed research issues related to the design of database management systems for geographic information systems and identified a long-term research agenda germane to the development of the next generation of geographic information systems. This paper summarizes the discussions that took place, and is an edited version of the longer report published as NCGIA Report 89-13.

Smith, T.R., Peng Gao and Gahinet, P. (1989) Asynchronous, iterative and parallel procedures for solving the weighted-region least cost path problem. Geographical Analysis 21: 147-168.

Based on research completed before the NCGIA award, this paper defines a family of procedures for finding least cost paths, using local, asynchronous, iterative and parallel processes. Although the procedures are guaranteed to terminate, it has so far been impossible to prove that they always terminate in admissible paths in the case of a two-dimensional triangulation. Extensive simulations have shown convergence to admissible paths in all cases examined.

6.2 Articles in Refereed Proceedings

Dorenbeck, C. and Egenhofer, M.J. (1991) Algebraic optimization of combined overlay operations. Proceedings, AutoCarto 10, Baltimore MD 6: 296-312.

The operations necessary to combine map layers are formalized with algebraic specifications. This shows that arithmetic operations upon discrete spatial subdivisions are reduced to a single, parametric overlay operation, the actual behavior of which is determined by a value operation which combines the non-spatial attributes of the individual cells of the corresponding layers. The novel approach is the application of these formalisms to find more efficient strategies for processing several overlay operations at an implementation-independent level. Two particular strategies are investigated: 1) the elimination of equivalent subexpressions to reduce the complexity of the overlay operation and 2) the integration of several overlay operations into a single one.

Egenhofer, M. and Frank, A.U. (1989) PANDA: an extensible DBMS supporting object-oriented software techniques. Proceedings, Database Systems in Office, Engineering and Science, Zurich. Informatik-Fachbenichte 204: 74-79.

The PANDA databases management system was designed for nonstandard applications which deal with spatial data. It supports an object-oriented program design with modularization, encapsulation, and reusability, and can be easily embedded into complex applications, such as spatial information systems or cartographic expert systems. Complex objects and their operations are defined. A layered structure on top of the programmer's interface provides object operations which include potentially complex consistency constraints.

Egenhofer, M. and W. Kuhn, (1991). Visualizing Spatial Query Results: The Limitations of SQL, in: E. Knuth and L. Wegner (eds.), *IFIP WG 2.6, 2nd Working Conference on Visual Database Systems*, Budapest, Hungary, *Elsevier*.

Several extensions to the relational database query language SQL have been proposed to serve as a spatial query language*; however, they do not sufficiently address how to visualize query results. This paper investigates the requirements for an* ad hoc *language describing the graphical presentation of spatial query results from the perspective of a geographic information system with frequent map output and assesses several spatial SQL extensions with respect to their treatment of the graphical presentation. It concludes that the SQL framework is inappropriate for this task at the user interface.*

Egenhofer, M., (1991). Beyond Query Languages for Geographic Databases, in: R. Demolombe, L. Fari nas del Cerro, and T. Imielinksi (eds.), *First International Workshop on Nonstandard Queries and Answers*, Vol. 1, Toulouse, France, pp. 131-137.

The application of traditional database query languages to geographic databases has been seriously hampered by the lack of functionality to formulate spatial queries and the shortage to represent spa-

tial query results appropriately. It is argued that the interaction between user and database has to be included into the language design to provide for a high-level, interactive database interface language.

Egenhofer, M.J., Frank, A.U. and Jackson, J.P. (1990) A topological data model for spatial databases. In A. Buchmann, O. Gunther, T.R. Smith and Y.-F. Wang, editors, Design and Implementation of Large Spatial Databases. Lecture Notes in Computer Science 409, Springer-Verlag, New York, 271-286.

There is a growing demand for engineering applications which need a sophisticated treatment of geometric properties. Implementations of Euclidean geometry, commonly used in current commercial GIS and CAD/CAM, are impeded by the finiteness of computers and their numbering systems. A spatial data model is proposed which is based upon the mathematical theory of simplices and simplicial complexes from combinatorial topology. It guarantees the preservation of topology under affine transformations. The implementation as a general spatial framework on top of an object- oriented DBMS is discussed.

Frank, A.U. (1991). Properties of geographic data: requirements for spatial access methods. In O. Gunther (ed.) Second International Symposium on Large Spatial Databases, SSD '91, Zurich. Lecture Notes in Computer Science 512.

Spatial access methods and the corresponding data structures, necessary to achieve the expected performance of Geographic Information Systems, are currently a prominent topic for research in spatial databases. Their performance is influenced by the properties of spatial data. This paper details the specific properties of spatial data in a GIS. It concentrates on GISs that store data describing objects with distinct identity and does exclude image and remote sensing databases, whose characteristics are very different, from consideration. We estimate approximate values for the parameters of some of the properties described and give measures for the size of GIS data. The results can be used for selection of a spatial access method for a GIS. They are also useful for the optimization of spatial access methods to respond to the specific requirements of a GIS.

Frank, A.U. and Barrera, R. (1990). The fieldtree: a data structure for GIS. In A. Buchmann, O. Gunther, T.R. Smith and Y.-F. Wang, editors, Design and Implementation of Large Spatial Databases. Lecture Notes in Computer Science 409, Springer-Verlag, New York: 29-44.

Efficient access methods, such as indices, are indispensable for quick answers to database queries. This article describes the fieldtree , a data structure that provides one such access method. The fieldtree has been designed for applications where range queries are predominant and spatial nesting and overlapping of objects are common. Besides their hierarchical organization of space, fieldtrees are characterized by three other features: they subdivide space regularly, spatial objects are never fragmented, and semantic information can be used to assign the location of a certain object in the tree.

Goodchild, M.F. (1990). Tiling Large Geographical Databases. In A. Buchmann, O. Gunther, T.R. Smith and Y.-F. Wang, editors, Design and Implementation of Large Spatial Databases. Lecture Notes in Computer Science 409, Springer-Verlag, New York: 137-146.

Geographical variation is infinitely complex, so the information coded in a spatial database can only approximate reality. The information will always be inadequate, in spatial resolution, thematic or geographical coverage. "Large" can be usefully defined as exceeding our current capacity to deliver. Traditional stores partition geographical data by theme and geographically. It is assumed that digital geographical databases will be largely archival, and will be similarly partitioned. A general model of a large archival store is presented. The properties of a generalized Morton key as a means of indexing tiles are analyzed, and its role in traditional systems of tile indexing is illustrated. For global databases, a tiling based on recursive subdivisions of the triangular faces of an octahedron using a rule of four is proposed. Earlier versions of this paper appeared in proceedings form (Goodchild, M.F. Optimal tiling for large cartographic databases. Proceedings, AutoCarto 9 (Baltimore, MD) pp. 444-51 (1989)).

6.3 Articles in Conference Proceedings

Barrera, R. M. Egenhofer, and A. Frank, (1992). A Robust Evaluation of Spatial Queries, Fifth International Symposium on Spatial Data Handling, Charleston, SC.

Evaluation of queries with requests for the aggregation of many detailed values in a database are of particular importance for Geographic Information Systems (GISs). They occur whenever a sum or a count for an area is requested and the individual data elements are stored. Geographic databases may keep versions of the same map with different levels of precision and these could be used to produce the answer more rapidly, perhaps with less precision. The more aggregated and less precise a representation is, the fewer instances are recorded and the less storage is occupied. For certain users and tasks, a result with less precision may be usable. For example, the display of an overview map may be sufficiently detailed based on the selection of a few significant objects. A trade-off between precision of the answer and reponse time asks for optimization. If one includes the factor of time available to perform a certain operation, it is possible to treat this as a trade-off between precision of the result and processing time: the more time available, the more precise one can determine the result. The requirements for such a system are (1) to perform incremental evaluations and (2) to assess how much a partial result deviates from the final, most precise result so that users can be information about the limitations of the answer.

Kuhn, W., (1992). Paradigms for GIS use. Proceedings of Fifth International Symposium on Spatial Data Handling, Charleston, South Carolina.

The conceptualization of the use of GIS is being considered here as a question of metaphor selection. The paper claims that understanding the use of a system as such, independently of specific tasks, is alway done in terms of some familiar domain of experience and is consequently metaphorical. In order to distinguish this understanding from the task-specific metaphors, the term “paradigm” is used. A paradigm can be conscious or can implicitly underlie the design and use of a GIS, but it always has profound psychological, economical and organizational consequences. Some relevant paradigms are analyzed and their current or potential role for GIS is discussed.

Smith, T.R., (1992). Towards a Logic-based Language for Modeling and Database Support in Spatio-temporal Domains, Proceedings of Fifth International Symposium on Spatial Data Handling, Charleston, South Carolina.

In this paper, the author justifies and defines a logic-based term definition language *(TDL) that facilitates the definition of interrelated domains of complex, spatio-temporal entities (which are either values or objects.) Expressions in TDL point to sets of procedures that are defined in a* function definition language *(FDL), and that provide semantics to the domains of elements defined in TDL. Such domains and procedures together form the core of a logic-based* modeling and database language *(MDBL) that may be tailored for specific domains of scientific investigation. TDL generalizes the concept of a data definition language, and, together with FDL, the concept of a data model; it subsumes the notions of type and class. We exemplify TDL by defining* effectively representable *domains of* Pointsets *and* Spatial_mappings *and show how standard spatial database systems (GIS) are expressible in TDL as special cases.*

6.4 Articles, Chapters and Monographs in Other Refereed Outlets

Buchmann, A., Gunther, O., Smith, T.R. and Wang, Y-F., editors (1990) Design and Implementation of Large Spatial Databases. Lecture Notes in Computer Science 409, Springer Verlag.

This volume is composed of the refereed papers presented at the Symposium on the Design and Implementation of Large Spatial Databases in Santa Barbara, July 1989, preceding the Initiative 5 specialist meeting.

Egenhofer, M. and J. Herring, (1991). High-Level Spatial Data Structures, in: D. Maguire, M. Goodchild, and D. Rhind (eds.), Geographical Information Systems, Volume 1: Principles and Applications, Longman, London, pp. 227-237.

Smith, T.R. and Je Yiang (1991) Knowledge-based approaches in GIS. In D.J. Maguire, M.F. Goodchild and D.W. Rhind (eds.) Geographical Information Systems: Principles and Applications. Longman Scientific and Technical, London 1: 413-425.

This chapter provides a framework for understanding the application of knowledge-based techniques in GIS. It is argued that full first order logic is a proper theory on which to base such techniques.

6.5 Articles in Other Outlets

Egenhofer, M. and Frank, A.U. (1989). Object-oriented modeling in GIS: inheritance and propagation. Proceedings, AutoCarto 9: 588-598.

The relational data model has proven to be too restrictive for applications with spatial data, such as GIS. The object-oriented approach seems to overcome some of the deficiencies. By incorporating the abstraction mechanisms of generalization and aggregation, the data model gets richer and more powerful than the relational model, and the application designer is given more and better

tools to model complex situations. Two methods for the derivation of properties are introduced: (1) inheritance describing properties and methods of subclasses in is-a hierarchies, and (2) propagation deriving properties in part-of hierarchies. While inheritance acts in a top-down fashion along the generalization hierarchy, propagation can derive values from parts to the aggregates (bottom-up).

Egenhofer, M.J. and Frank, A.U. (1990). Object-oriented software engineering considerations for future GIS. Proceedings, IGIS Symposium.

Currently, experts in software engineering promote an object-oriented approach for the implementation of large software systems. This advanced technique, based upon the definition of object types in combination with the corresponding operations in a modular fashion, is characterized by clear coding which can be maintained easily. Powerful object-oriented abstraction mechanisms, such as classification, generalization, and aggregation, are the framework for an object-oriented model. Specific programming languages have been designed to allow programmers to implement software systems pursuing an object-oriented design. Future GIS will benefit from object-oriented abilities to concisely define complex problems.

Egenhofer, M.J. and J.R. Herring, (1991). A framework for the definition of topological relationships and an algebraic approach to spatial reasoning within this framework. Report 91-7, National Center for Geographic Information and Analysis, Santa Barbara CA.

A new theory of binary topological relationships between n-dimensional spatial objects is presented. Unlike previous approaches, it provides a complete coverage, i.e., any possible constellation between two spatial objects can be described by exactly one of the relationships defined. The formalism is based upon fundamental concepts of algebraic topology and set theory. Spatial regions are modeled as point-sets and the binary topological relationships are then defined in terms of the intersections of the boundaries and interiors of two point-sets. Sixteen potential relationships are identified by considering empty and non-empty intersections. Prototypes are shown for the eight relationships that actually exist between two point-sets embedded in a two-dimensional space. More detailed relationships as refinements of these eight relationships are identified by considering other criteria, such as the number of the individual segments of the four intersections or their dimensions. A simple, exemplary system is described that performs reasoning about the spatial relationships between members of a set of spatial objects. The main problem of interest is to make sound and complete inferences about the set of all spatial relationships that hold between the objects, given prior information about a subset of the relationships. The spatial inferences are formalized within the framework of relation algebra and procedurally implemented in terms of constraint satisfaction procedures. Although the approach is very general, the particular example employs the "complete" set of topological relationships described above. In particular, a relation algebra for these topological relations is developed, and a computational implementation of this algebra is described. Systems with such reasoning capabilities appear to have many applications in geographical analysis, and could be usefully incorporated into GIS and related systems.

Frank, A.U., (1993). Design of cartographic databases. In J.-C. Muller (ed.) Advances in Cartography. Elsevier.

Before an intelligent discussion of cartographic databases can start, we have to clarify the necessary terminology: "What is a cartographic database and how is this notion related to other similar terms, GIS in particular?" We stress the importance of the database concept and detail some technical aspects. The contents of a map can be conceptualized in different forms, from a purely graphical viewpoint to a highly structured collection of data, each set of concepts carrying with it its appropriate set of operations. After establishing the differences between GIS and cartographic data base, we then go on to explore potential relations between a GIS and a cartographic database. Of special interest is how a GIS and a cartographic database for the same region are related and how one can benefit from the other by establishing links between multiple representations of the same real objects. Technically speaking, the cartographic database is a 'materialized view' of the GIS; therefore methods like triggers and active database concepts need to be explored for their suitability. The major problems in designing cartographic data structures are a lack of understanding of the structure of maps and the process that produces them, and the lack of model for map reading. The linkage between data structures and map structure is stressed and research topics outlined.

Frank, A.U. (1989). Requirements for database management systems for large spatial databases. In R.E. Dahlberg, J.D. McLaughlin, B.J. Niemann Jr., editors, Developments in Land Information Management, Institute for Land Information, Washington.

The paper presents a list of functions a database management system should provide for GIS, and some quantitative estimates are given on the size of a typical graphical retrieval. The main part of the paper shows a possible layered architecture for a Geo DBMS, including the physical storage level with clustering and buffer management, a layer to protect data, access methods using key values, spatial access (supported by physical clusters), data modeling concepts and corresponding operations, support for abstract data types and object-oriented programming, methods to deal with consistency constraints, and query languages. This architecture has been used for the construction of the PANDA database system.

Hudson, D. (1990). Autonomous view-stages: materialized support for view update propagation. PhD Thesis, University of Maine, Orono.

We offer here a description of a database query optimization technique called the method of reduced autonomous view-stages. This method provides a compromise solution to problems found in current query optimization strategies which occur when the only query evaluation choices are (1) to bear the cost of repeatedly recomputing the same query results or (2) to maintain a fully evaluated materialized view which is expensive to update. The technique presented here is a specific engineering solution only. It is not universally effective or all views under all query conditions, but is intended to be implemented for a particular view only after comparative cost analyses with other methods. The contributions of this work are (1) in the introduction of the novel approach of view- stages; (2) in providing an algebraic and procedural interpretation of the concept called the factorability of a view such that it is autonomously computable with respect to updates; (3) the analysis of many aspects of factorability; and (4) an algorithm that provides a fast approximation of a minimal cost view-stage for comparison with alternative optimization methods.

Lanter, D.P., (1992). "GEOLINEUS: Data Management and Flowcharting for ARC/INFO", Tech-

nical Software Series S-92-2.

Geolineus Data Mangement and Flowcharting for ARC/INFO" Abstract: Geolineus is a new way of working with GIS that displays the structure of GIS databases graphically and keeps track of the GIS data and applications as they are built. Geolineus lets users store meta-data about what source data represent and automatically tracks the lineage of GIS layers derived from these sources. Structured meta-data serves as the basis for Geolineus's data management, documentation, and quality control/quality assurance capabilities.

Lanter, D.P. (1990). Lineage in GIS: the problem and a solution. Report 90-6, National Center for Geographic Information and Analysis, Santa Barbara CA.

The first paper in this pair focuses attention on a fundamental geographic structure: the GIS application. Lineage documentation specifies an application's source data, transformations, and input/output specifications. Such information is inherently causal, communicating the theory embodied in a GIS application and the meaning of its product. A number of techniques for automating lineage information are examined. None are found to be capable of documenting data lineage. The second paper, "Design of a Lineage-Based Meta-database for GIS", presents the conceptual design of a meta-database system for documenting data sources and GIS transformations applied to derive cartographic products. Artificial intelligence techniques of semantic networks are used to organize input-output relationships between map layers and frames to organize lineage attributes characterizing source, intermediate, and product layers. An illustrative example indicates that a lineage meta-database enables GIS users to engage in source assessment throughout their analysis of spatial data sets.

Smith, T.R. and Gao Peng (1990). Experimental performance evaluations of spatial access methods. Proceedings of 4th International Symposium on Spatial Data Handling, Zurich: 991-1002.

Many spatial access methods (SAMs) have been developed in response to the growing needs of GIS and non-standard DBMSs. No single SAM as yet has been proven to outperform all others in every aspect of performance. Past experimental evaluations have been carried out in a relatively ad hoc manner and the various experimental results cannot be used to make comparisons between the SAMs. We have conducted a comprehensive experimental performance evaluation on several representative SAMs in terns of a systematic experimental design. We employed a set of parameters characterizing the distribution of objects in the database, including object frequency, object density, object size distribution ratio, and object spatial distribution as a basis for modeling file structure performance. The response surface models were in general significantly nonlinear, with wide variation between the performance of the various SAMs in relation to the parameters.

Smith, T.R., Ramakrishnan, R. and Voisard, A. (1991). An object-based data model and a deductive language for spatio-temporal database applications. Proceedings of the GOODS Workshop on Spatial Databases, Capri, Italy, Springer-Verlag.

We discuss data models and languages in relation to computational environments in which data access and tools for analysis are integrated in order to support data-intensive and numerically in-

tensive modeling activities in a class of scientific database systems. We describe a simple model of scientific activity in order to motivate the development of a new conceptual data model for complex, spatio- temporal objects and a deductive, object-oriented language to support the definition and manipulation of such objects. We provide three examples of the application of both the model and the language.

Trivedi, N. and Smith, T.R. (1991). A conceptual framework for integrated metadata management in very large spatial databases. Report 91-2, National Center for Geographic Information and Analysis, Santa Barbara CA.

A conceptual framework for integrated metadata management in large spatial databases is described. The primary function of this framework is to allow definition, location and control of metalevel information about the underlying database. The framework provides for a set of core metadata components and allows for addition of any auxiliary metadata that the user might want to define. The framework would support feature based retrieval as well as interactive browsing of metadata. The emphasis is on flexibility, extensibility and ease-of-use. The goal is integrated management of all kinds of metadata. The report gives an overview of semantic modeling of spatial data followed by a conceptual model for metadata. The basic tenet behind the conceptual model is classifying the database entities of interest into data, process and environment entities. Corresponding to this, the metadata consists of metadata, metaprocess and metaenvironment entities. We then propose a forms mechanism to manage metadata. A set of basic operations for manipulating forms and catalogs is described. We present a case-study of metadata in a conventional GIS environment. This is supported by the preliminary version of the schema for the Condor Database Project at the University of California at Santa Barbara.

Weber, C.R. (1991). A screen-reflection algorithm for calculating buffer distances in raster images. Technical Papers, ACSM/ASPRS Annual Convention, Baltimore MD 2: 325-332.

A new algorithm employing RAM-resident graphics calls is presented which significantly decreases the currently reported computation time for growth of simple Euclidean buffers in raster forma geographic information systems. Buffer zones are created onscreen through graphics calls for each selected feature-pixel within a scan-line file. The resulting screen display is then compared with the original file and buffer pixel values from the display are placed within a new file. Completion time or the algorithm is less than 50algorithms currently in use.