GLOBAL INFORMATION SYSTEM ISSUES

Iris Kameny
The RAND Corporation
1700 Main Street
Santa Monica, CA 90406

A Global Information System (GIS) is an umbrella system that presents a global schema of the data available in its underlying, component, autonomous information systems and provides its users with the capability to ask queries of the global schema and to receive integrated answers. (Note that developing a global schema requires the resolution of conflicts between underlying schemas and is not merely the result of concatenating the underlying schemas.) At this time, there are no commercially available GISs though there are experimental research systems such as:CCA's Multibase, Honeywell's DTS, and Unisys's Mermaid.

The GIS differs from distributed data management systems (DDMSs), such as IBM's R*, Tandem Nonstop, Ingres/Star and Sybase in that its underlying information systems may be managed by different DBMSs using different data models. Like several DDMSs, the underlying information systems of the GIS may run on different hardware bases utilizing different operating systems. Although all commercial DDMSs will support retrieval and data updates, most GIS experiments have been limited to retrieval only services, only a few have addressed the update issue. Currently, most GISs require that the component systems communicate through established networks and standard communication protocols.

The importance of GISs is being realized as the information revolution proceeds and what were previously autonomous, stand-alone DBMSs or information systems are now connected by networks and available to users through workstations. The ability to develop an enterprise-wide view of data and access to information through that view is recognized as a critical base on which to build decision aids and distributed applications. In fact, the act of building an integrated view of enterprise-wide data is expected to yield benefits in providing a better understanding of how the enterprise operates and its future data needs.

The major research areas confronting GIS developers are (1) development of global schema; (2) query processing and data access; (3) security; (4) recovery from failure; (5) integration of the user workstation with the GIS; (6) support for multi-media objects. The first four research areas are most important because a viable GIS depends on their solution. Areas 5 and 6 are becoming more important as technology advances.

(1) The development of global schema addresses the heart of the GIS system. A discussion of major issues in integrating schema can be found in Batini et al's article titled "A Comparative Analysis of Methodologies for Database Schema Integration" (ACM Computing Surveys December 1986) in which they describe four steps for developing a global schema: preintegration, comparison, conformance, and merging and restructuring, and the need for tools to support these activities. The GIS global schema describes all of the data in the component systems; all component schemas and user views are mapped to it. The process of understanding the enterprise objects and their relationships and mapping the component schemas to the global representation can be very complex involving the resolution of many differences among component schemas such as in names, types and structures. There may also be a need for specialized views of the integrated schema to simplify usage or to protect data. A logically centralized data dictionary/directory (DD/D) is used to represent the global schema and related information (e.g., information about objects, their relationships, and mappings into the component schemas; view definitions; and location and replication information).

Ongoing research is addressing tools for data modeling, schema integration, and DD/D representation, development and use. Future research needs include: management of global schema development including maintenance of the global schema when the underlying schemas change; and development
of a rich constraint language (for example, using rules and metadata).

(2) Query processing and data access includes language and protocols for remote access, query optimization, and concurrency control, and is a very active research area. Query processing involves decomposing the query into multiple queries, planning how to acquire the answer in an optimal way, translating the decomposed query into proper target language queries, establishing network connections to carry out the process, monitoring the process, adjusting the plan, and composing the final answer to return to the user.

Many research efforts are addressing query optimization including exploration of performance enhancement through parallelism. Several database standards efforts are engaged in defining a common set of operations addressable through a standard language to be used by all DBMSs. This would enable the development of GISs and could provide the basis for distributed concurrency control for transactions and updates. Currently, GISs have no easy way to control component DBMSs in order to assure concurrency and most do not support updates or transactions (though several research efforts are addressing this problem). (For example, a GIS query will be decomposed into several queries which are sent to component DBMSs without control over concurrency therefore permitting possible data inconsistencies.) Another area of research important to DDMSs and worth examining for GISs is monitoring of user needs and patterns to perform load balancing by redistributing or replicating data. By definition, a GIS could not do this automatically (since the components DBMSs are autonomous) but through integrated management could suggest changes in the component systems which would result in better performance.

(3) Security includes security of data, authorization and authentication of users, and accountability--all of which are important issues in the development and use of a GIS. Authorization management may be completely decentralized with each user having a userid for each component system with authentication and permissions determined by the component DBAs and/or Security Officers. In this case the GIS becomes a front-end service in connecting users to the underlying systems each of which has responsibility for authenticating users and controlling access to data. This approach may violate system transparency if users need to login to each component system requiring access. Authorization may be centralized by each user having a GIS-wide userid and by the GIS authenticating the user and determining permissions based on GIS maintained data security views. More research is needed in this area to better understand authorization issues and choices, accountability issues (e.g., maintaining audit trails for security and billing), and the handling of data and schema at multiple levels of classification (government) or sensitivity (industry).

(4) Recovery from failure addresses failures in user workstation nodes, component DBMS nodes, GIS nodes; and network partitionings. The main research issues are to explore the recovery techniques needed for different kinds of system failures, and examining the trade-offs between making the GIS more reliable and available by the use of replicated data and multiple network connections against complicating query optimization. Other research issues include how to determine failures and how/if/when to return partial answers when part of the underlying system is unavailable;

(5) Integration of the user workstation with the GIS is an area that has not received much attention. The research issues include how to integrate a local data schema with the GIS global schema for individual use, how to integrate workstation databases into the global schema, presentation of GIS multi-media objects, and, in general, how to share functionality between the workstation and the GIS (since the workstation may be like a component DBMS node in some respects but different in others).

(6) Support for multi-media objects is an issue that is beginning to receive research attention. Several GISs access multi-media objects as files through their descriptions in component databases. Research topics include location of multi-media objects by content through specialized operators, temporary storage needs for very large multi-media objects, and network protocols for transporting very large multi-media objects efficiently.