ADVANCED GEOGRAPHIC INFORMATION SYSTEMS

Claudia Bauzer Medeiros
Institute of Computing, University of Campinas, Campinas, Brazil

Keywords: database integration, environmental planning applications, geographic databases, geoinformatics, geographic information systems, GIS science, information technology, remote sensing data, transdisciplinary research, urban planning applications, Web information management

Contents

Summary

Geographic information systems (GIS) are large sets of integrated software modules, developed around a database management system, which manage, analyze, and display geographic data. Being pieces of software, they are fated to a speedy decay, given the rate of progress in information technology. The advanced system of today is outdated the very next day. How, then, can "advanced" systems be presented? One of the challenges faced by this theme is how to discuss geographic software systems avoiding this quick obsolescence.

The approach taken here is to give an overview of these systems as a combination of four factors: people, data, technology, and software. "Advanced" are the kinds of data managed, which are collected taking advantage of contemporary technological progress. "Advanced" is the kind of information system described, which incorporates state of the art research in computer science, in terms of database management, algorithms to process the data, and to derive relevant information thereof, as well as mechanisms to present and let users interact with these data. "Advanced," furthermore, are the transdisciplinary applications developed using these systems, which allow the information being managed to be put to new uses. But—and here lies the basis of the real advance—"advanced" are the people who develop these pieces of software and technologies, who collect these data and use the applications. The "advanced" adjective applied to people implies that they are progressing intellectually and technically at a very fast rate, and growing more demanding. Advanced geographic information systems, and the applications built using them, are thus software that manage distinct

kinds of geographic data and whose goal is to empower people, facilitating their education and enabling them to better understand their world. This software can thereby help people take appropriate decisions on how to live in, improve and preserve their world.

1. Introduction: The Information Society and Geographic Information Systems

Information technology is heralded as the key to the future. The Internet has transformed the world into a global village, in more senses than one. Just as the Industrial Revolution of the nineteenth century modified the standard of living of mankind through machine (horse)power, the Information Revolution is transforming their world through information (byte)power. Computer networks have eliminated the notion of distance as an impediment for cooperation among people, and are profoundly changing the meaning of social interactions. Virtual enterprises, virtual communities and digital government are examples of concepts that popped up with this new revolution.

From a computer connected to "the" network, anyone can run an immense amount of applications using data that are spread all over the world. The 320 million documents on the Web in 1998 grew to 800 million in 1999, amounting to over fifteen terabytes of data, with an expected growth rate of 1,000 percent every couple of years. Approximately 372 million people in twenty-one countries connected to the Internet from their homes in February 2001, jumping to 379 million in March of the same year. This interlinked and interwoven world has become a global information system, whose users are the planet's inhabitants. Grid computing, where machines distributed all over the world cooperate in processing large simulations, is yet another facet of this scenario. The dynamics of such a system are beyond our understanding.

What is an information system? Broadly speaking, it is a computer-based set of software modules that interact with a database, processing and transforming its data into information that is meaningful to some set of users. The information generated by this system may be fed back into its database, or into some other system's database. If this definition is extended to the global network, the world is the database, the software is everywhere, the hardware that runs and feeds the system can be anything and the users are everyone.

Several factors in this global information system add up to a very unstable configuration, for whose management there is no foreseeable solution. First, anybody can access anything anywhere: the emergent concept of ubiquitous computing. Second, data are continuously fed into this system by people, software and devices, under different formats, scales, quality standards, entry rates, and volume. These data have to be made sense of, "digested," and processed by the system. They are, furthermore, ephemeral: the average lifetime of an Internet Web page is two months, and some pages may last for only a few hours. Third, hardware and devices can connect off and on to this system, feeding it data and processing the available information. Fourth, this system is composed of an astronomical number of software modules developed by countless people in multiple, distributed versions. Finally, all the components of the system are elusive: people, hardware, and software are nomadic.

There are no means to manage this, unless some partitioning is performed: of *data*, *software*, *people*, and *hardware*. Data-centered approaches try to organize this world database by means of defining data collection, filtering, preprocessing, and display procedures based on factors such as kinds of data sources or user access patterns. Software-centered approaches organize the modules into libraries of software components. People-centered approaches are concerned with developing interfaces to this global system to help users connect to and navigate within it, and with establishing new mechanisms to support virtual co-operation. Hardware and infrastructure-based procedures concentrate on performance, reliability, and availability.

But what has all this discussion to do with Geographic Information Systems (GIS)? GIS are first and foremost information systems, whose goal is to handle geographic data. They are, like any information system, subject to (and victims of) this new era of worldwide information management. Since much of the information needed for decision-making in our daily life is spatially referenced, geographic data comprise a large portion of the data fed daily into the global network. These data are voluminous and massively collected all over the world for a multitude of geographic applications. *Geographic applications* consist of software in which the geographic location is a determinant factor. These applications may concern individuals (e.g. to help choosing one's itinerary to work or deciding where to build a house). They may also involve communities (e.g. planning sanitation infrastructure, monitoring urban traffic conditions, mapping land use, managing forested land, estimating crop production, assessing water quality, protecting wildlife, and many kinds of global change monitoring).

To understand advanced GIS, one must keep the global framework in mind. It is furthermore essential to differentiate between the core of an information system and the applications that are developed using this system. An information system is a database-centered software package that serves as a basis for implementing applications, and is often bought off-the-shelf. Applications are additional software layers developed "on top of" (or using) this package, customized for specific user profiles. For instance, office information systems are software packages that allow office workers to interchange files and establish and manage common work procedures. Office applications are software developed using office information systems, but which are tailored to different organizational culture and needs.

Figure 1 gives a schematic view of these concepts. It shows an information system as software composed of three layers: database management system (DBMS), data processing modules, and programing interface. Neither applications nor databases are at the core of an information system. Applications are developed using these layers. Databases are sets of files stored "underneath" and managed by the information system. These files may be all in one place or distributed on a network. Applications and database are specific to different organizations or user communities. The DBMS is the only software that can access and update the database, thereby ensuring a minimum of integrity and reliability.

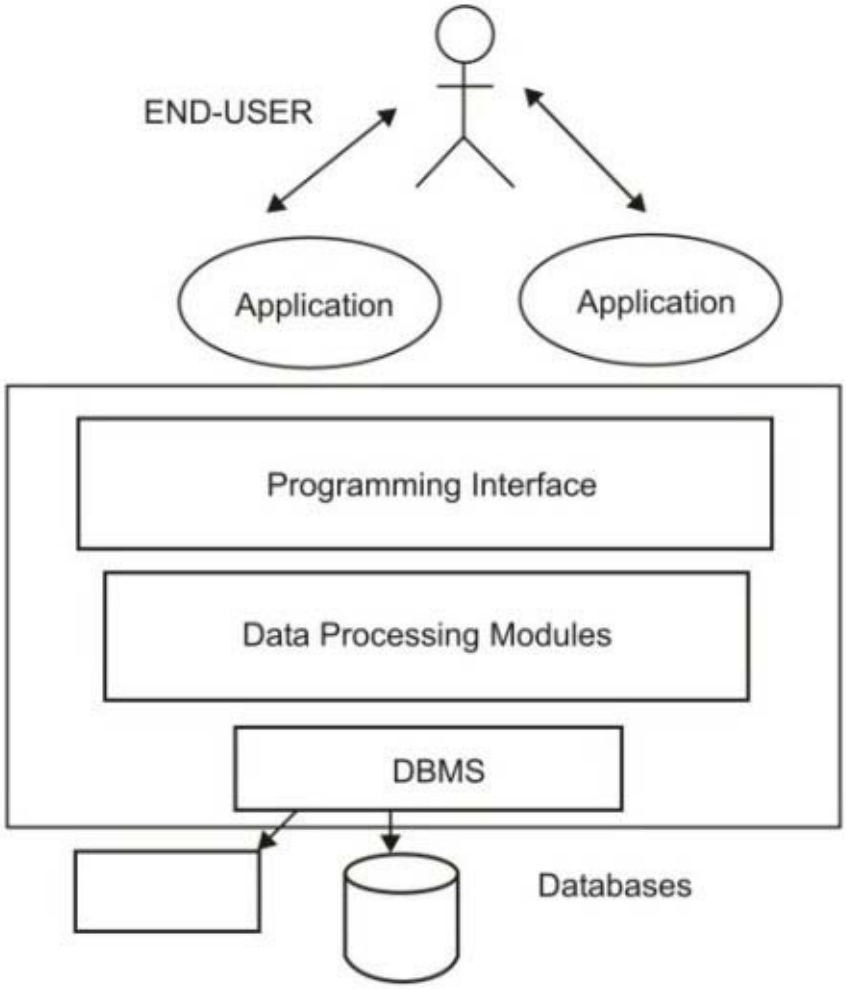

Figure 1. Schematic presentation of a general purpose core information system, in a three-level diagram, comprising programing interface, data processing software modules, and database management system.

A geographic application is thus any piece of software that helps users solve some problem involving geographic space; the GIS is the software infrastructure necessary to develop the application. Four kinds of factors contribute to the construction of geographic information systems: data, people, software, and technology, examined in the subsequent sections. Thus, the value of the information that is derived from and produced by a geographic application depends on:

- the quality of the data in the geographic database
- the quality of the geoprocessing technology used
- the quality of the GIS and application software
- the expertise of the people who will analyze the information generated by the applications and use it to produce knowledge.

Unfortunately, these factors are by nature intermingled: one cannot talk about people without characterizing the applications they demand; one cannot discuss applications without situating them in the context of GIS software and associated technology; this, in turn, requires discussing the data, which prompts considerations on their use, and

therefore on people – the users. There is no established, canonical, way of ordering these factors when discussing geographic systems. The presentation that follows favors and is colored by a computer science approach.

Section 2 covers aspects of the data managed by geographic systems. These data are generated and processed by devices and technology discussed in Section 3. Section 4 examines the software modules that compose a GIS, differentiating between the system and the applications. Section 5 considers the people involved in making and using GIS and its applications. Section 6 presents some advanced geographic applications under the framework of data, technology, software, and people. Section 7 sums up the paper, indicating some future trends; using this framework to let the reader envisage other trends.

2. Geographic Data: What Are They, and How To Integrate Them?

Geographic applications are motivated by people's needs, but they are data-centric, data-intensive, and data-driven. Each one of these characteristics defines requirements and peculiarities of these applications, and of the systems on which they run. Data "centric-ness" has to do with the fact that geographic data are at the core of these applications, which are concerned with how entities behave and interact in the world. In this context, understanding and modeling data are key issues. Data "intensive-ness" is linked to geographic data volume and continuous evolution. Applications and systems have to provide means of adequately managing these large data volumes, for different time and space scales. Data "driven-ness" signals the causal dependency amongst data collection, application execution, and result reliability. This, in turn, has repercussions on user interaction facilities, performance, and integrity constraints.

These issues have to do not only with the intricacies of modeling, sampling, and storing, which are computer-related. They also are a direct consequence of the fact that each person sees the world differently and needs to gather distinct information from it. Thus, there are as many views of the world as there are people in it. One may even consider that there are non-human perspectives as well (e.g. to adequately model the behavior of a given animal species in its habitat, one may need to incorporate the animal's perception of the world into the model).

Different views and needs result in distinct data files for the same geographic region. It may be necessary to combine these files to obtain an integrated perspective of the region. This process is called data integration and is perhaps the most complicated problem in the management of geographic databases. This section deals with the question of how to model, store, and integrate data in order to try to accommodate all these views.

2.1. Defining, Modeling, and Storing Geographic Data

What are geographic data, after all? They are basically data collected or deduced from anything in the world, using the globe as the basic canvas onto which different scenarios—the applications—are painted. Anything on earth, animate or inanimate, can be considered as a geographic data source. One may tend to consider that only visible or

tangible entities are sources of geographic data. However, derived data (e.g. spatial relationships and trends) may not be visible or directly measurable, but also constitute geographic data. One cannot touch these data, but one lives under their influence.

The need to store and handle geographic data in a computer forces people to move away from real world problems and to become concerned with low-level implementation details. Unfortunately, one cannot escape from these details at some point, if only because it is impossible to grasp every fact about everything.

The process used to create a database goes through three stages: modeling the entities of interest in the world; modeling the database to store data captured about these entities; and creating the database, "filling it" with these data. Thus, the term "geographic data model" has two distinct meanings in GIS literature. One concerns how the world is modeled and the other how database experts specify the geographic database. Both concepts are based on determining the geographic entities of interest and the relationships among them.

In the first perspective, geographic entities in the world may be modeled either as continuous functions varying along some surface (this is called the "field model"), or as discrete entities with identifiable boundaries and identity (this is called the "object model"). Object and field models are eventually translated into some database model. The latter is a formalism that helps specify how to build the database files. These files may be defined in terms of tables (relational data model), or interrelated, nested, sets of records and functions to apply to them (the object-oriented data model). Finally, once the database is specified, data can be stored in it.

Modeling (world entities or a database) is a complex activity, since it requires interpretation and sampling. A complicated issue involves the object versus field decision. A first approximation classifies man-made artifacts as objects (e.g. cities, roads, bridges) and natural phenomena as fields (e.g. temperature or elevation). This is misleading, since the distinction between fields and objects is scale- and application-sensitive. A mountain is an object identifiable by name, but its relief is a continuous field.

This discussion is induced by the fact that we can precisely measure and model relationships among man-made artifacts, but not among natural phenomena. For instance, we can describe a building's perimeter and height, or the topology of highway networks, and have constructed instruments to record these measurements. On the other hand, perimeter and height do not often apply to natural entities (e.g. what is a forest's height?) Precise modeling and measurement capability versus approximations thereof have given origin to two ways of storing geographic data: "vector" and "matricial" (also called "raster"). In the GIS literature, these representations are sometimes also called model (i.e. vector model, raster model), aggrandizing the terminology problem. The vector representation stores geographic data in terms of points, lines, and polygons. The matricial representation relies on storing these data as matrices, frequently images, where each matrix element corresponds to a pixel (picture element).

Once fields and objects are used to model the world, it is time to model the database, defining its files, records, and record fields (the attributes). Geographic database records contain two kinds of attributes: descriptive and spatial. Consider for example a river and a highway. Typical descriptive attributes for highways would be number, width, and construction material; river descriptive attributes might be river name, average depth, water temperature, and chemical composition. The spatial attributes are stored under vector or matricial representations. Under the vector representation, polygonal lines can represent the geometrical and topological properties of river or highway. Under the matricial representation, these real world entities appear as clusters of pixels in an image.

To allow computation of spatial relationships (e.g. distance), geographic records must be "geo-referenced" (i.e. associated with their location on the Earth). Geo-referencing is part of the preprocessing operations needed before the data are stored in the database. In the vector representation, geo-referencing is achieved by associating coordinates to points (and thus to lines and polygons). In the matricial representation, geo-referencing associates coordinate values to pixels.

Data modeling, sampling, gathering, and preprocessing procedures are of foremost importance for any kind of geographic application, and have a direct impact on the reliability and quality of the applications that use these data. Data maintenance procedures, such as update policies and scheduling, also influence applications. If updates are not performed often enough, information is valueless; if updates are performed too often, maintenance costs will rise.

Geo-referenced data values are a result of sampling and estimates—the database on the highway or the river contains in fact data on some representative points thereof, so there is always an element of error. This may not be so important in the case of man-made artifacts, where educated guesses often allow a representation closer to the truth. However, in the case of natural phenomena, this varies according to the goals of who or what is preprocessing the data, and the assumptions that underlie the model that has been chosen to mathematically model these phenomena. This, in turn, has a profound impact on the subsequent uses of these data and of their integration with other data.

2.2. Gathering and Integrating Geographic Data

Geographic applications require merging physical, social, political, economic, environmental, and engineering information. Geo-referenced data files are heterogeneous by nature. They differ from each other in terms of how their data were captured and preprocessed, at which time and space scales, for which goals and quality control procedures. Allied to their volume and permanent evolution, this transforms the creation, integration, and maintenance of geographic databases into a major challenge. Integration of heterogeneous data, regardless of whether they are geographic or not, is a deterrent for any application. It is an open research problem in computer science and has been so since the early 1970s.

The term "database integration" means a process through which distinct data files are reorganized and articulated into a database according to a unifying principle. It does not

necessarily mean that all data are stored together or that they have to be modified. It just means that they are perceived—by people, software, or hardware—as being part of a whole. Database integration may be achieved physically or virtually. Virtual integration is obtained by developing software to provide the integrated view, without changing the original data. Physical integration requires that data be modified. In the world of heterogeneous, networked data, integration is increasingly virtual, and the files of the integrated database are distributed.

In order to integrate data, one must know their provenance or genealogy. The questions to be asked of any geo-referenced data set are: who/what collected and inserted the data into the files, when this insertion happened, and what is the data's validity and volatility? Another set of questions concerns how the data were collected: devices used, sampling procedures, updating frequency, interpolation, and approximation functions. First and foremost, however, applications are sensitive to geographic factors (e.g. the region where data were collected, at which scale, and when), the kind of data being managed, and the purpose for which they were collected.

All of this information (who, what, where, when, how) is considered to be metadata, or data about data. There are several metadata standards proposed for geo-referenced data, but there is no consensus on which factors to consider. Metadata have acquired importance in the context of the Web. They are one of the means used for integrating data from multiple sources. As an example, consider integrating two files containing the average daily temperature of some cities. Associated metadata serve to check data compatibility. For instance, metadata will register whether temperature is stored in degrees Celsius or Fahrenheit, and thus indicate if there is a need for conversion. Furthermore, these data can only be compared if they are considered within the time and location frames in which they were collected.

Metadata are also used to help systems interact with each other: the so-called "interoperability" problem. Metadata standards are a data-centric approach to integration and interoperability. A companion people-centric approach to these problems is the use of ontologies.

In computer science, an ontology is a description of a set of objects and their relationships. This includes the vocabulary for referring to the terms in a subject area, and rules that describe how these terms can (or cannot) be related to each other. Typically, ontologies are stored in large textual files using a graph structure. This graph represents a taxonomic hierarchy among the vocabulary terms. Terms can moreover be linked to each other, indicating distinct kinds of semantic relationships. Additional files contain rules that describe situations in which the vocabulary may be used, and which may help derive relationships among the terms. A thesaurus is an example of a simple ontology. Ontologies are used in artificial intelligence to support communication among people and/or software modules. In database research, especially when the Web is concerned, ontologies are used to establish correspondence among data records, thereby helping to integrate data and to find information in the Web. In the temperature example, an ontology might indicate that Celsius and Fahrenheit are temperature measurement units. Ontologies in GIS are proposed, for instance, to define relationships among distinct conceptualizations of space. For instance, when geographic data on a

given region are collected by different people with distinct goals (e.g., a team working on biodiversity issues and another on ore prospection), strategies such as matching or alignment of ontologies may be used to help integrate the resulting data files.

Even when the data integration process may seem clear-cut, the procedures for integrating geographic data sources may well vary according to the applications that will use these data. Two databases that were created as a result of integrating exactly the same set of data sources may differ because of the integration criteria. Therefore, in spite of the abundance of geographic data already available, it is often the case that an application demands new data collection procedures for the same kind of geographic entity: data collection is ultimately prompted by application needs. An urban traffic application does not need to consider individual parcels in a city, but these are the most important data units from a tax application perspective. Neither application will need to regard temperature, which is essential in urban air pollution studies.

Thus, the questions to be answered in order to create a database for an application are: which data to collect, where, when, how, and how often? Since data collection is very expensive, a good alternative is to reuse existing files. In this case, questions raised include: are the needed data already available, and if so, where and in which format? These questions fall into the computer science fields of data mining, information retrieval, and data discovery. Equally important questions involve which data to integrate and how? Ideally, collection and integration procedures should cater to all applications that might ever want to use these data. Since these needs are unpredictable, this is not a feasible recommendation. These concerns lead naturally into the next section that will discuss the technology of capturing, processing, and storing data into a geographic database.

-
-
-

Bibliography

Berners-Lee, T., Hendler,J., Lassila,O. 2001. *The Semantic Web*. Scientific American, May 17. [The paper that launched the idea of the Semantic Web, which is the basis of countless implementations of GIS on the Internet, stressing the importance of interoperability and the role of semantics in this]

Brooks, F. P. 1995. *The Mythical Man-month, Anniversary edition: Essays on Software Engineering.* New York, Addison Wesley. 175 pp. [Continuously in print since 1975, this classic book provides a valuable historical survey of how and why large software projects fail. Though not directly related to GIS, it should be read by anyone involved in developing large GIS applications.]

Couclelis, H. 1992. People Manipulate Objects (But Cultivate Fields)—Beyond the Raster-Vector Debate in GIS. *Springer-Verlag Lecture Notes in Computer Science*, No. 639, pp 65–77. [A clear discussion on the differences between field and object views, and their computer representations.]

Crawford, D. (ed.) 2001. The Next 1000 Years. *Communication of the ACM (CACM)*, Vol. 44, No. 3. pp. 27–148. [*CACM* is the flagship journal of the Association for Computing Machinery. This special issue contains fifty-four papers from renowned experts in computer science and education on future forecasts for science and technology.]

El Masri, R.; Navathe, R. 1999. *Fundamentals of Database Systems*. 3rd edn. New York, Addison Wesley. 960 pp. [A well-structured, undergraduate level text on how to design and construct databases, with a good presentation on the internals of database management systems.]

Elmargarmid, A. K.; Rusinkiewicz, M.; Sheth, A. (eds.) 1998. *Management Heterogeneous and Autonomous Database Systems*. San Diego, Morgan Kaufmann. 413pp. [A set of papers from experts in the domain of database integration covering current approaches to the problem.]

Goodchild, M. F.; Jeansoulin, R. (eds.) 1998. *Data Quality in Geographic Information: From Error to Uncertainty*. Paris, Editions Hermes. 192 pp. [Set of papers discussing the several dimensions of geographic data quality.]

Goodchild, M. F.; Parks, B. O.; Steyaert, L. T. (eds.) 1993. *Environmental Modeling with GIS.* New York, Oxford University Press. 488 pp. [A collection of texts giving examples on several problems concerning environmental modeling and implementation using geographic information systems.]

Gregg, S. M.; Leinhardt, G. 1994. Mapping out geography: an example of epistemology and education. *Review of Educational Research*, Vol. 64, No. 2, pp 311–61. [Very good paper on how to teach through geography.]

Guenther, O. 1998. *Environmental Information Systems*. Berlin, Springer-Verlag. 244 pp. [Book that covers several issues in development environmental systems and applications, with examples of existing systems.]

Hassler, S. (ed.) 2001. Always On: Living in a Networked World. *IEEE Spectrum*, Vol. 38, No. 1. pp. 14–123. [Every January, this journal of the Institute of Electrical and Electronics Engineers publishes several short technical papers on the main developments of the previous year in science and technology, and future perspectives. This 2001 issue is dedicated to Internet innovations, emphasizing the role of GPS and wireless communication in several applications.]

Jones, C. 1997. *Geographical Information Systems and Computer Cartography*. Essex, England, Addison Wesley Longman Limited. 319 pp. [A good textbook on several technological aspects of GIS internals, geographic data capture and electronic cartography.]

Longley, P. A.; Goodchild, M. F.; Maguire, D. J.; Rhind, D. W. (eds.) 1999. *Geographical Information Systems: Principles, Techniques, Applications and Management*. 2nd edn. New York, John Wiley. 1058 pp. [This classic two-volume book provides a comprehensive overview of GIS theory and applications, written by experts in the field.]

Monmonier, M. 1991. *How to Lie with Maps*. Chicago, University of Chicago Press. 176 pp. [A very good text on the dangers of taking cartographic renditions at face value.]

Morain, S.; Baros, S. L. (eds.) 1996. *Raster Imagery in Geographic Information Systems*. Santa Fe: OnWord Press. 495 pp. [Didactic text on the technology to capture and process different types of image data for GIS applications, with many examples. Though slightly out of date, it contains a good explanation of the basic theory behind this technology.]

UNESCO—IIEP (eds.) 2001. *Quality and learning: perspectives from development co-operation. Report on the meeting of the International Working Group on Education (IWGE).* Paris, UNESCO. 127 pp. [Good overview of the problems in managing education in the era of information technology, which must be considered in the development of GIS educational software.]

Want, R.; Schilit, B. (eds.) 2001. Location-Aware Computing. *IEEE Computer*, Vol. 34, No. 8. [Special issue containing papers on technological and societal aspects of applications that are sensitive to users' location.]

Biographical Sketch

Dr Claudia Bauzer Medeiros is a full professor of Computer Science at the Institute of Computing of the Universidade Estadual de Campinas (UNICAMP), Brazil. She received her Electronic Engineering degree in 1976 and her M.Sc. degree in Informatics in 1979 from Pontificia Universidade Catolica do Rio de Janeiro, Brazil. Her Ph.D. degree was awarded by the University of Waterloo, Canada, in 1985, in the area of database systems.

She is the head of the database research group in UNICAMP, and leads projects on different aspects of geospatial databases, including query processing, versions, spatial decision support systems, and the management of integrity constraints and temporal data. Other research interests include image databases and documentation management. One of her major concerns is the design and development of applications centered on geo-referenced databases, aggregating interdisciplinary teams, for environmental and urban planning. She has co-ordinated several large research and development projects around GIS technology in Brazil, and was the senior database consultant for the development of many GIS applications in the domains of urban planning.

She was for six years the Editor in Chief of the *Journal of the Brazilian Computer Science Society*, the main scientific publication of the Brazilian Computer Society (SBC), and is on the editorial board of *GeoInformatica* and the *VLDB Journal*. She has been a keynote speaker or participated in program committees of major database and GIS conferences, having been the program chair of the Seventh ACM GIS Symposium. She is one of the co-founders of the Brazilian Special Interest Group on Databases, and a member of ACM, IEEE, and SBC.

She was awarded the 1996 and the 2001 Academic Merit Prize of the University of Campinas, and the 2000 Newton Faller Prize of the Brazilian Computer Society.