METHODOLOGY FOR ASSOCIATING A CONVENTIONAL DATABASE TO A GEOGRAPHICAL INFORMATION SYSTEM: THE SAGA/UFRJ CASE STUDY

OSMAR MOREIRA DE OLIVEIRA
LUÍS FERNANDO BARBOSA DE ALMEIDA
JORGE XAVIER DA SILVA

UFRJ (Universidade Federal do Rio de Janeiro), Laboratório de Geoprocessamento, Rio de Janeiro, RJ, Brazil
Ph#:(021)270-6186 - Fax:(021)598-3280 - E-mail: XAVIER@ufrj.bitnet & XAVIER@ibm.nce.ufrj.br (Internet).

Abstract. Many aspects of our environment can be integrated as data recorded in a database. By environment it is understood not only the natural one but also the environment built by the man. In this sense we refer both to raw data concerning the rain forest as well as to data relating land ownership information. These data are very important for statistical analysis, but they usually lack of a fundamental point: the territorial expression. Thus, there must exist some type of linkage, relating graphical information (usually a map that represents a region of interest georeferenced to any cartographic system) to conventional information (represented by information records containing ancillary data in tabular form: vegetation species of a forest, parcel owner data, etc.).

This paper presents a methodology for integrating such diverse data in the context of a raster Geographical Information System (GIS), the SAGA/UFRJ [8,9,10], developed at the Laboratory of Geoprocessing of UFRJ. This methodology covers all the steps of the work, from the initial planning phase, passing by the logical project of the conventional database, to the logical project of the integration of the data to the GIS.

Another important aspect discussed in this paper refers to the user characterization, showing that even the most experienced users need the help from a geoprocessing specialist (analyst) in some phase of their work. This is fundamental to reduce excessive expectatives and turns easier the execution of a work.

1.INTRODUCTION

Geographic data are commonly characterized as having two major components [1,2]: the phenomenon being reported and the spatial location of the phenomenon. A database is a collection of related data (facts that can be recorded and that have implicit meaning) [5]. These two definitions are very important to characterize the role developed by a GIS: its ability to integrate georeferenced data, which includes operations like spatial search and overlay [1,2]. Besides, a digital map cannot contain all the information (attributes) related to spatial location in its graphical data structure, but a link must exist that relates a spatial location with a given database record(s). In the same manner, it should have a way to add georeferencing to the conventional database records.

This paper addresses these problems in a straightforward way:

- presenting the raster GIS SAGA/UFRJ, showing its data structure and modeling;
- defining logical projects of a database;
- proposing a specific model to relate the two previous defined systems (the model really implemented in the SAGA/UFRJ), and
- commenting the action users must take when facing a project to solve using such an integrated system.

2.THE SAGA/UFRJ

The SAGA/UFRJ is a raster GIS designed to run on low-cost equipment. As data entry it uses A-4 scanned pieces of the original document (normally a map), joined later in digital form inside the system. As output three options are available: VGA PC-monitors, color matrix and color ink-jet printers.

Someone could argue what is the advantage of such a system, and why a vetorial GIS was not used directly? There are many reasons, between them [1]:

- it is a simple data structure;
- one of the most important operation on a GIS is overlay, which is facilitated if the maps are on a raster format;
- high spatial variability is efficiently represented;
- image processing operates at raster level;
- even for typical vector-operations, like network analysis, we are developing raster solutions, with very good preliminary results [3];
- it is possible to analyze relevant topological

relationships as we will show below;

- the raster represents the data in a more realistic (although less aesthetically pleasing) than its vector counterpart. Zooming operations in raster show the data at most basic unit: the pixel. In opposition, a vector system zoom maintains the topology at any scale, although some approximation and manipulation, is always executed but under a sort of hidden form;
- new data compression techniques allow more efficient storage of matrices.

The SAGA/UFRJ is basically composed of three modules: MONTAGEM, TRAÇAVET and SAD.

The first one (MONTAGEM) join up - and export - a large map (bit map TIFF format) with smaller imported A-4 pieces, recorded in non-compressed, 256 gray-tone TIFF format. For doing this it operates over each small map, converting it to the metric system, refining its borders, rotating it and transforming it to a bit map. The basic unit of the SAGA/UFRJ is one element of the matrix related to a spatial square (resolution). The spatial attributes (reduced to the three topological and fundamental geometric elements: point, line and area) are entered as aggregates.

The second module (TRAÇAVET) operates upon the output of the module MONTAGEM and provides an efficient storage of the map. This is done in two phases named neo-vectorization and superimposition of the classes over the raster.

For each homogeneous set of classes in each map (for example, hidrography, streets network, etc.) is created a file (*.vet) in which the "vectors" are added. There is also a special field in the general header that will be responsible for the association with the conventional database. Then, the neo-vector files are superimposed over the raster, generating new legends that compound the new raster file (*.rst).

The features (represented by the neo-vector files) are superimposed over the raster, with each file (and all its "vectors") associated with only one legend. This restriction demands an extra caution when defining the homogeneous set of features that will be "vectorized", since from the raster (without a database link which is the alphanumeric database code) it is impossible to recover one particular feature contained in a neo-vector file.

The third module, SAD, performs analytical functions over the raster maps. Basically it is directioned to environmental studies allowing the following analysis: environmental signature (from one spatial position retrieve all the spatial attributes locally associated), environmental evaluation (weighted overlay) and monitoring - reporting the changes observed in time.

The scope of this paper calls for a more detailed description of the data entry process, mainly the process operated by the TRAÇAVET module, specifically the process of inputting cartographic data which is detailed below:

1. output of the module MONTAGEM (bit map refined);
2. previous organization of the legends needed for the map;
3. creation of the neo-vector file by the recognition of the features associated to each legend;
4. creation of the new raster map by the superimposition of each neo-vector file of every legend;

Although those modules are very important to environmental analyses they lack some specific functions to deal with land ownership data and other non-spatial attribute. A very important function is the ability to link cartographic data to conventional databases (like dBASE and others) and, in addition to some analytical functions related to land planning. In this regard, as a final consideration about the SAGA/UFRJ, two geoprocessing modules can be mentioned: a) TRAJETÓRIA [3], destined to define minima and maxima paths along friction surfaces. The friction factor may be associated with time, cost, distance, etc.. As soon as a cadastral map is available with its linkage to a conventional database, this module assists very interesting analyses for transport routing, emergency dispatching and other similar network applications; b) SAGA-BD [10], to be explained below.

3.DATABASE DESIGN FUNDAMENTALS

Database design [4,5,7,6] has three major phases: the conceptual, the logical and the physical designs. The conceptual design is characterized by:

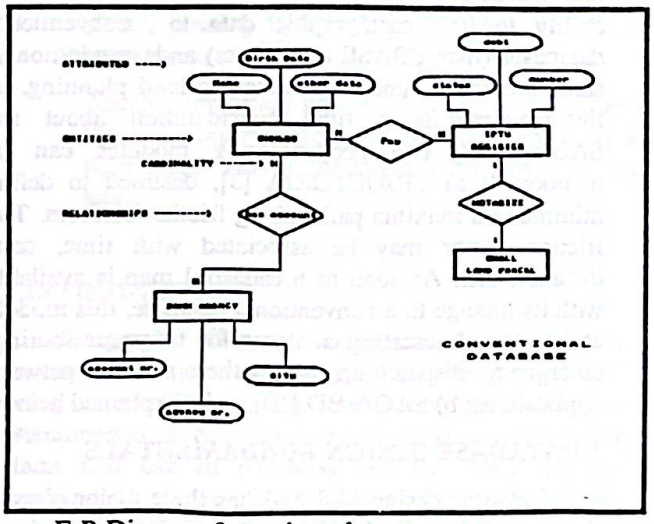
- analyses of the data needs;
- determination of the data interrelationships;
- independence of the DBMS (database management system).

Once a conceptual design has been executed, the logical and physical designs must be created for the specific DBMS (database management system) by an expert.

The model assumed in this work for the conceptual design is the Entity-Relationship (E/R) Model. The basic concepts behind the E/R Model are:

- entity: object in the real world with an independent existence (for example THE parcel at the address: 3, Fountain Square);

- entity set: set of entities of the same type (for example the blocks (pieces of land containing many parcels and bounded by streets) in a given city may be defined as the entity set BLOCK.
- attributes: particular properties used to describe each entity, for example the entity PARCEL can be described by the attributes address, owner, area, built area, perimeter and municipal register.
- key attribute: allows to distinguish each entity belonging to a set of entities of the same type. For example the address identifies unambiguously the parcel;
- relationship: it is an association connecting entities. Other important concept refers to the cardinality of a relationship. Cardinality specifies the number of elements of an entity set which can be related to each element of another set of entities by the relationship. The figure 1 below shows a simplified



E-R Diagram for cadastral purposes.

Figure 1: Simplified E-R Diagram for Cadastral Purposes

4.ASSOCIATING A CONVENTIONAL DATABASE TO THE SAGA/UFRJ: SOME CONCEPTS

When linking a conventional database to the SAGA/UFRJ, the main problem is to create the key attribute(s) that can describe both the conventional database and its cartographic representation map in the GIS. Defined such a key, it will guide the conceptual design of the database (as explained in the previous section) - which will orient the lower levels designs (logical and physical). At this point the link file structure must be created. The figure 2 illustrates this fundamental step of addition of a new entity to the pre-existing E/R Diagram.

To alter the conceptual design of the original conventional database **(without this conceptual design, nothing could be done for completing the association with the SAGA/UFRJ)** some guidelines must be followed as indicated below.

4.1.CONCEPTUAL DESIGN ALTERATION

1.identify in the E/R Model the entity(ies) with geographical representation and its key attribute(s);

2.create a new entity (which we call GEOGRAPHICAL REPRESENTATION) and a new relationship (HAVE) with cardinality ratio of 1:1 (remember that this new entity is chosen because it has a key attribute, i.e. an unambiguous identifier), linking the entity identified to its geographical representation mentioned;

3.the new entity, GEOGRAPHICAL REPRESENTATION, has a unique attribute: geographical key;

4.spreading of the alteration to the entire conceptual design of the conventional database.

When those attributes are already defined, and the database analyst is provided with the corresponding orientation, it is necessary to create the link file to the GIS using the module ACESSO-BD (sub-module of the SAGA-BD module). This file is structured in a matricial form in which, for each pixel, we insert a link code. The structure of this file is created at this time with the same dimension of the map.

It should be stressed again that without the database link it is impossible to recover individual neo-vectors from a neo-vector file, once they are superimposed over the raster. One of the implicit advantages of creating such a link is to allow this type of individual recovery.

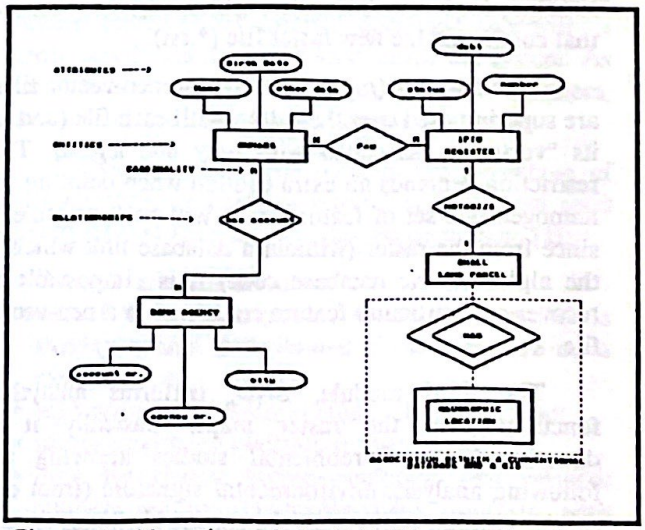


Figure 2: Alteration of the E/R Diagram for linkage purposes

The module CONSULTA-BD (another sub-module of the SAGA-BD module) allows cross queries between the conventional database and the digital maps stored in the GIS. Both the ACESSO-BD and the CONSULTA-BD will be explained ahead.

In order to allow a correct association of the conventional database above delineated to the cartographic database the process previously described in section 2 (described as part of the SAGA/UFRJ) must be done with same cautions:

After step 2 (legends organization) it should be chosen the data aggregation level associated with the problem (integration unit to be referred to the database). With this done, the features associated to each legend must be recognized on the bases of this new aggregation level). Each set of homogeneous features will generate a neo-vector file, allowing the identification of it individual components.

In this step, each neo-vector receives a title which can be used as an entity key (carrying spatial representation), already defined in the E/R Model as shown in figure 2. This title could then be used later as a way of recovering a particular feature inside a neo-vector file, as will be defined in the next section.

After step 4 the legends just created must be associated with the link file already created.

4.2.THE ACESSO-BD FUNDAMENTALS

As mentioned before, the ACESSO-BD is a sub-module of the SAGA-BD, interface responsible for associating a conventional database to the SAGA/UFRJ. It has, as main function, the generation of the link file from the neo-vector files. The procedure for doing this, performed by the ACESSO-BD, is detailed below:

1. Neo-vector files must be chosen so that they can be used for doing the wished association (remember that, at this point, the key attribute - related to the new entity GEOGRAPHICAL REPRESENTATION - is already chosen). Those neo-vector files must correspond to the respective entities contained in the database.

2. Each neo-vector file receives a unique number (recorded in the field DATABASE CODE of the general header of each neo-vector file, *.vet - see section 2) named SEGMENT. Besides, each neo-vector inside that file receives a code (recorded at a temporary contact file) which is the sum of the order of this neo-vector inside that file (OFFSET) and the DATABASE CODE (SEGMENT).

3. This code, named LINK CODE hereafter, is associated with one specific neo-vector file, and will be superimposed to the link file (remember again that this link file in matricial form, similarly a SAGA/UFRJ analytical raster file, but must not be confused with it).

4. As mentioned before, with the link file it is created another file, the contact file, that is simply an ASCII file relating each individual neo-vector, through its title, with the link code (locational keys) just created. Each neo-vector title represents possible key attributes of the entities in the relationship HAVE (see figure 2). Thus the temporary contact file can be used to stablish wished relationship inside the conventional database and be discarded afterwards.

4.3.THE CONSULTA-BD FUNDAMENTALS

This module, like the ACESSO-BD, is also a sub-module of the SAGA-BD module. It is responsible for supporting cross queries between a digital map (with its associated legends) and a conventional database.

There are two possibilities of using these queries. As mentioned before, both require that the conceptual design had been altered and propagated through the design of the conventional database. The first possibility aims was to isolate the SAGA/UFRJ from the database system used, performing the communication between them via a special text file (that must be called #LISTA.TXT). In this way, a user can do queries in their own system and get spatial results in the GIS (in our case the SAGA/UFRJ) simply writing the tabular queries results down to that special text file. This represents a great advantage when compared with the GIS softwares that have proprietary databases. The second possibility is represented as a sub-program of this module - CONSULTA-BD - that accepts one DBF (dBase file standard) file representing a single table with as many fields as necessary and, of course, the key attribute associated with the new entity GEOGRAPHICAL REPRESENTATION (of the conceptual altered design). With these concepts in mind, the next two sections will describe the two possible types of queries (DATABASE -> MAP and MAP -> DATABASE).

4.3.1.QUERIES FROM THE DATABASE TO THE MAP

This query can be done basically knowing the name of the link file. The results of the tabular query done (through the procedures described in the previous section) are stored in temporary file #LISTA.TXT. This file directly presents the link codes associated with that feature represented in the database and there respective link file names.

There are two ways of relating them: directly through the points (pixels) of the link file - searching the link (matrix) file for the spatial positions associated with those codes and plotting them directly on the map, or through the neo-vectors (set of pixels) identified in the link file. In this file we have two references: the link code by itself (remember that it is equal to the sum of the SEGMENT and the OFFSET, where the SEGMENT identifies the neo-vector file as a whole and the OFFSET indicates the relative position of the neo-vector inside the file) and the name of the neo-vector file that contains the searched neo-vector. Then, once identified, it is plotted on the map. This last method although seeming slower than the first one, in reality is more efficient, mainly with large maps where the processing of the link (matrix) file is slow.

4.3.2.QUERIES FROM THE MAP TO THE DATABASE

This association is done directly from the link (matrix) file to the database. The user chooses one or more features from the map in the screen. The pixels chosen are immediately identified in the link (matrix) file, giving the corresponding link codes for the database. Similarly these codes are written down to the special file #LISTA.TXT which is read by the database system, providing the wanted records directly.

5.CONCLUSION

The integration of a conventional database to the SAGA/UFRJ (GIS) has proved to be a strong and useful tool to the environmental analysis and, more, as well, to urban and rural studies. This integration, like the SAGA/UFRJ project, is based on low-cost technologies.

The link procedures were developed as independent modules, communicating through files associated to the georeferenced features on the map and vice-versa, thus allowing queries based on conventional database technology. Hence, it is a generic methodology that can be used by any commercial database system based on PCs. This fact is of relevance since many research institutes and governmental institutions already have their own conventional database designed to meet their needs.

REFERENCES

[1]ARONOFF S., *Geographic Information Systems: A Management Perspective*, Canada, WDL, 1989.

[2]BURROUGH P. A., *Principles of Geographical Information Systems for Land Resources Assessment*, USA, Clarendon Press, 1987.

[3]CAMPOS A. C. S., OLIVEIRA O. M., XAVIER-DA-SILVA J., Trajetória - Um novo módulo para o SAGA/UFRJ, *Anais da IV Conferência Latino Americana sobre Sistemas de Informação Geográfica - 2º Simpósio Brasileiro de Geoprocessamento*, 733-747, São Paulo - BR, 1993.

[4]DATE C. J., *An Introduction to Database Systems*, USA, Addison-Wesley, 1981.

[5]ELMASRI R., NAVATHE S. B., *Fundamentals of Database Systems*, USA, The Benjamim/Cummings, 1989.

[6]FURTADO A. L., SANTOS C. S., *Organização de Banco de Dados*, BR, Ed. Campus, 1986.

[7]KORTH H. F., SILBERSCHATZ A., *Database System Concepts*, USA, McGraw-Hill, 1991.

[8]XAVIER-DA-SILA J., CARVALHO-FILHO L. M., Sistemas de Informação Geográfica: uma proposta metodológica, *Anais da IV Conferência Latino Americana sobre Sistemas de Informação Geográfica - 2º Simpósio Brasileiro de Geoprocessamento*, São Paulo, BR, 1993.

[9]XAVIER-DA-SILVA J., SOUZA M. J. L., *Análise Ambiental*, Rio de Janeiro - BR , Editora UFRJ, 1988.

[10]XAVIER-DA-SILVA J., et alli, Um Banco de Dados Ambientais para a Amazônia, *Revista Brasileira de Geografia*, v.53 no.3, BR, IBGE, jul/set - 1991.