Full Integration of Geodata in GIS

A. SCHAUMBERGER

Socrates-Erasmus Summer School, Brno 2006

1. Introduction

The main approach of GIS is to describe parts of the world in more or less complexity, especially in its spatial context. In generally, information systems help us to manage all the information which is collected and processed. In the case of a Geographical Information System (GIS) this information is combined with spatial orientation according to the position in the real world. An information system and a GI-System respectively, consists of several components like hardware, software, methods, people, and data which have to work together for overcoming the tasks of different subjects. GIS integrates many disciplines and creates the links between them, the technology for solving geographical problems, and the science behind this technology.

The data are the most important part of a GIS – they are like the fuel for running a machine. This paper concerns to the several aspects of data and their integration in a GIS. Geographical data or geodata are usually very heterogenic concerning their sources, types, structures, and qualities. The integration of different kinds of data in a homogenous system like GIS is often a big challenge, but one of the most important preconditions for modelling, analyzing, and visualisation of spatial phenomena.

In order to avoid errors in results of applications in GIS it is necessary to follow on rules for correct data handling. GIS software supports you in many cases with a huge number of functions which can be easily misused. You have to understand the procedures behind the functions and their effects on your geodata. This work explains the most important facts concerning geodata integration which must be considered. It should also help you to create correct datasets for proper decision making processes of your stakeholders.

Additionally to the theoretical knowledge which is needed for useful data handling some examples illustrate the practical approach of data integration. These examples refer to vector data as well as raster datasets and combinations of them, respectively.

2. Abstracting the real world in geodata

The objects we find in the real world can be represented in a ‘binary environment’ for computing them in several ways. This transformation process is determined by the level of abstraction. Real world objects are very complex and cannot be represented completely in data models. For that you have to reduce the information as much as you need it for your application. More information which is not necessary for a certain task causes more problems concerning storage and effectiveness. Therefore it is very important to think carefully about your needs and definitions of abstraction levels. The transfer of objects from the real world to our databases or GIS environments follows concepts which should strictly be taken into account.

2.1. Sources of geodata

All GIS project needs data which have to be collected. The source of collection can be the real world or already existing data. You can distinguish between primary and secondary data. Primary data are measurements or observation results of entities in the real world. The remote sensing data like orthophotos or satellite images are examples of such an observation. GPS measurements, surveying data, or digital gathering of natural inventories by field experts are also primary data sources.

Secondary data are the results of operations on existing data. Most of these data are analogue, for example paper maps. The maps are often the base of digitising work in order to extract certain objects. Another important secondary dataset is the Digital Elevation Model which is created by the evaluation of maps.

2.2. Types of geodata

In a geodatabase you may find a lot of different types of data. Generally, geodata consists of more or less attributes in combination with geometric information. The kind of geometry depends on the aim of your application. It can be based on the vector data model or on the raster data model. If you want to describe objects with discrete boundaries the vector data model is more suitable. In the case of continuous data which cover completely a certain area the raster data model should be used.

The vector data are built with points as their atomic unit. Two points can be connected to a line feature. Some lines can be aggregated to a polyline and if this polyline has its starting

and ending point on the same position you get a polygon. These are the three different features which are the base of all vector oriented datasets. The advantage of this data structure is the efficient use of storage and a lot of geoprocessing functions which can be applied on it. Otherwise the algorithms are more complicated and often need long computing times.

The raster data model is very simple compared to that of the vector data. It consists of rows and columns with mostly equal extended cells. Each cell contains a value which describes a situation for the area covered by the cell. For example, the Digital Elevation Model is a raster dataset with the elevation value in each cell. The algorithms for analysing raster data are simple and easy to use. Unfortunately, raster datasets need a lot of storage capacity.

2.3. Quality of geodata

The quality of data is very important especially you need them not only for visualisation but also for analysing. Many operations would run incorrectly if the quality aspect of the geodata is neglected. The quality of data concerns the accuracy and precision of creation, the maintenance of topology over the whole data management process, and the correct use of data types and functions applied to it.

The topology is necessary for the consistence of a dataset. For example, if you have ‘Spaghetti data’ you cannot distinguish between polylines and polygons or define exact crosses of lines. These problems, for example, make network analyses impossible.

Many datasets are used in different software systems and therefore they have to be consistent for conversation approaches. GIS software reacts more or less sensitive on bad data quality. If the data will be distributed you have to make sure that no information are lost or changed by the integration process.

Actually, the care about data quality should be always in your mind if you deal with geodata in any way. The maintenance of a consistent dataset takes a lot of effort and is also expensive but only high quality data promise good results of GIS analyses.

2.4. Data about Data – Metadata

In many cases the data have a lot of implicit information which is absolutely necessary for correct integration in GIS. This information is an essential part of the geodata and is called

metadata. For example, if you get data from unknown source without these metadata you cannot find out the meanings of abbreviated attribute names or the applied projection and much other important information.

The metadata are an indicator for the quality of geodata and have four roles (CAR, 2006c):

- Availability – Data needed to determine the sets of data that exist for a geographic location
- Fittness for use – Data needed to determine if a set of data meets a specific need.
- Access – Data needed to acquire an identified set of data.
- Transfer – Data needed to process and use a set of data.

For a quick overview the metadata should inform you about the name of the dataset, the developer, the covered geographic area, the currency of data and any existing restrictions on data. The general quality of data should be described and inform the user how good the data are (suitability, accuracy, completeness, consistency). A very important part of this subject is the information about spatial reference which means the coordinate type, the projection, the datum, and the conversions parameters. Some additional information about format, online availability or costs is also very useful (CAR, 2006c).

2.5. Data model concept

The transformation of real world objects to geodata is an analytical approach. The complexity of phenomena is reduced by splitting in different data layers according to the data types which are needed.

The process of transformation is structured in three main steps. First you need to define a conceptual model which allocates the real world information to objects or to smooth continuous spatial information. The next step is the definition of data models concerning to the conceptual model. A set of discrete objects, their attributes and relations or a continuous smooth field is defined. The third step represents the data models either in a vector or in a raster data model (CAR, 2006b).

The topographic situation of a certain area, for example, is stored into a raster data model and used as a Digital Elevation Model. The rivers or streets on this area are another layer of lines, the parcels are polygons and the villages are points. The data type depends on the

scale of your application. Cities can be modelled as points in small scale applications or as polygons for a larger scale visualisation.

The physical data model describes the way of storage and organisation of logical data model definitions which are mentioned above. In most GI-Systems the geodata are organised in relational models. This means, that geometric information is linked to the attribute information through unique identifiers. Both types of data are usually stored in separated tables. This structure can be implemented in a geodatabase within a Database Management System (DBMS), e.g. MS SQL-Server for the ESRI ArcSDE Geodatabase. But you can find this structure also in a file oriented system like the ESRI Shapefile (dbf for attributes and shp for the spatial data).

In the following you find examples for using geodata sources in different systems which are not described in detail but give an idea of meaningful use of such data.

3. Integration of Geodata

The starting point of all GIS analyses and visualisation is the availability of geodata. In order to generate new knowledge on base of several geodata you have to combine these data sources. The combination of geodata in GIS is the main approach you have to deal with for running GIS applications. You can put your collected data together and analyse or visualize them in a traditional way or you integrate these data logically (see *figure 1*).

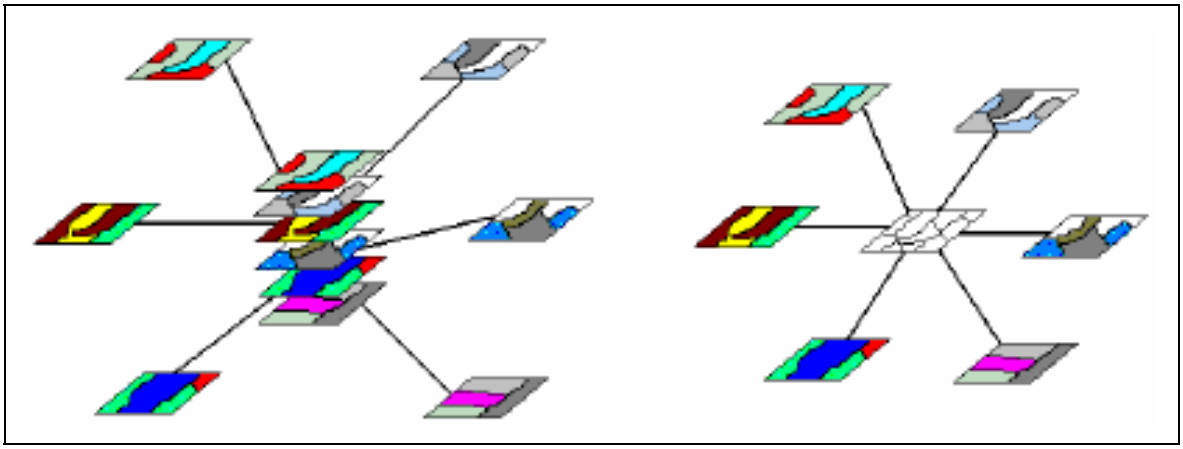

Figure 1: **Difference between collecting and integrating of geodata (KOLEJKA, 2006a)**

Today there are a lot of geodata available, but in very different formats, scales, projections, etc. The management and especially the integration of these heterogeneous data is a

synthetic process of overlaying several themes and often a big challenge. The application of a Digital Landscape Model could be a suitable solution and its concept is presented below.

3.1. Digital Landscape Model (DLM)

Digital Landscape Model is at least 3D-4D computer generated scheme of a segment of the earth's landscape sphere providing in a simplified, but integrated form its basic structural and dynamic features. It results in a new type of a geodata base. The entities of a real word area are classified in four main types of geographic data layers (KOLJEKA, 2006c):

- Natural background (with homogenous natural landscape units as reference areas)
- Products of human impacts (with parcels and/or subparcels as reference areas)
- Human and social interests (development limits in parcels and/or subparcels as reference units)
- Digital Elevation Model (carrying skeleton)

These four aspects of a landscape can be additionally extended by expert systems for multidimensional approaches. Examples of DLM applications are soil loss modelling, risk management, run off modelling, land suitability assessment, landscape planning, landscape historical research, etc. The models are not only for scientific use but also for presentation, explanation, and description. Visualisations in 3D, for example, are very impressive and provide an intuitive access to complex information.

The process of DLM construction follows certain rules. First of all you collect the geodata you need for your DLM. The collection need often work on data digitising, georeferencing, or format unification. After that the geodata have to be integrated by some integration methods you selected. These methods can be manual, semiautomatic (on-screen integration), or automatic (clustering and classification techniques). The next step is the selection of a reference layer as the most reliable geodataset. The order of integrated layers goes hierarchically from background (reference layer) to indicator information (KOLEJKA, 2006c).

The attributes for identification of integrated geodata in DLMs are built by vectors of expressions of the involved data layers, for example *v = (K, H, T, G)*. That means that attribute *v* consist of a combination of coded information of the different DLM layers (*figure 2*).

The individual feature description has to be put in consequent attribute table columns. So you have to have separate columns for e.g. climate, soils, geology, slope, humidity, etc. for analytic utilisation (KOLEJKA, 2006c).

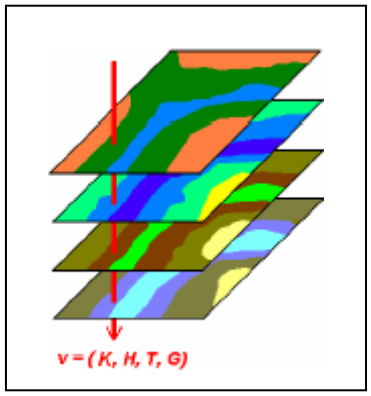

Figure 2: **Use of data overlay in DLMs**

3.2. Land Use Changes Modelling

The modelling of changes are simulations of processes in the real world, which can be executed statically or dynamically. The data models behind are mostly continuous field models, network models or tessellated space models (polygons, TINs, Grids). A simple evolution model uses the same rule for all grids without interactions between neighbouring grid cells. It is a simple prediction model for one attribute. The local dynamic models are more complex. It describes dynamic interactions of a number of local parameters. The coupled dynamics, single-system model takes additionally into account the interactions between neighbouring grid cells. The coupled dynamics, multiple-system model use multiple models for different grid cells together with interactions between neighbouring gird cells. The most complex model is one with dynamically changing structure. There are multiple models for different grid cells with neighbouring interactions and a local model may be replaced with a new one (some properties inherited). (ROOSAARE, 2006a).

The time is a key factor in all change modelling approaches. For example, climatology, hydrology, and also human studies deal with changes during certain time periods. The changes of land use which imply human activities can be modelled by several tools.

IDRISI software supports these tasks with some functions according to MARKOV algorithms. The current version of IDRISI, the Andes Edition, provides the Land Change Modeller for Ecological Sustainability which need well prepared input data. This tool helps you to find changes in the past as a basis for prediction. You are able to generate a model for these changes and predict land use changes in attaching human interventions.

The Land Change Modeller has a huge number of setting options which help you to precise your results iteratively. The validation, of course, is a very important part of all modelling tasks. The tool offers several functions (e.g. VALIDATE, GROSSTAB, ROC) to apply the validation which compares the predicted with the actual change.

The IDRISI software tool needs a lot of geodata for the calculation and is therefore a problem of data integration. The data are often incomparable and have problems concerning scale, generalisation, resolution, feature definitions and much data are not explicitly spatial. The results of land use changes models have to be harmonised with expert knowledge (ROOSAARE, 2006a).

3.3. Climatic Data Extrapolation

Weather data are measurements on certain stations – they are point representations of meteorological phenomena. In GIS these data need to be processed in a full area coverage dataset for analysing. In the case of temperature you have to take into account the dependence of elevation. The extrapolation of temperature data cannot be executed without the separation of this dependence. Therefore the integration of a Digital Elevation Model is needed.

The Regression between temperature and elevation calculates the values which are needed for applying this dependence on the Digital Elevation Model. The residuals are substituted with a second equation which takes also into account the proportional isolation and the variability of the topography concerning slope and aspect. The coefficients are the results of empirical studies. Both full area coverage datasets are finally added with map calculator. The result is the extrapolation of temperature for a certain area in dependence of the relief.

IDRISI offers the full functionality to carry out all the necessary tasks. The statistical functions you can apply on your raster datasets are completely included. Furthermore all actions you have to do can be modelled by the Macro modeller, stored and reused for other projects.

4. Discussion

Geodata are the most important part of GIS. The integration of data in respect to their differences is one of the main tasks in GIS with a wide range of aspects. In many projects

you need geodata from several sources, in different formats, scales, resolutions, qualities, spatial references, etc. Standardisation, for example, is an important issue in many fields of technology. It minimizes the effort of integration and combination of heterogeneous subjects and it is also the basis for applying interoperability. This is a very important aspect in GIS, too. Especially you have to deal with geodata on distributed platforms like Geo-Portals on the Internet.

There are many initiatives to implement a common standard of GIS mainly concerning the geodata. International groups like ISO, INSPIRE, or OGC tried to do this in the past and they were successful in some parts. The improvement of standardisation is the big challenge for the future. The positive impact of this approach on geodata integration has to be considered and should not be underestimated.

The integration of geodata shows a wide variety of implementations and depends on the systems which are used. There are no common rules or schemes how to do this integration. It differs from system to system, from model to model, even from project to project. The Socrates Erasmus Summer School “Full Integration of Geodata in GIS” presented several examples how to deal with this task in different ways.

5. Source material

All references are presentations at the Socrates-Erasmus Summer School: Full Integration of Geodata in GIS, Brno, 21.05.06 – 02.06.06.

CAR, A. (2006a): GIS Data Sources – GIS Introduction.

CAR, A. (2006b): GIS Data Sources – Sources of Geodata.

CAR, A. (2006c): GIS Data Sources – Metadata.

CAR, A. (2006d): GIS Data Sources – GIS & Spatial Decision Support Systems.

GUSZLEV, A. (2006): Cartographic visualisation in GIS.

KARYDAS, Ch. (2006a): Geo-Data Input and Conversation.

KARYDAS, Ch. (2006b): Uncertainty and errors in GIS.

KLIMANEK, M. and J. KOLEJKA (2006): Stationary climatic data extrapolation using digital terrain model.

KOLEJKA, J. (2006a): Full Integration of Geodata in GIS – Core Ideas of Socrates Erasmus Summer School.

KOLEJKA, J. (2006b): Logical data integration into digital landscape model 1.

KOLEJKA, J. (2006c): Logical data integration into digital landscape model 2.

MIKITA, T. (2006): Multimedia Presentations of GIS Outputs.

ROOSAARE, J. (2006a): Modelling in GIS.

ROOSAARE, J. (2006b): Geospatial Analysis in GIS.

VICENS, L. (2006): Multidimensional Presentations in GIS.

VRANKA, P. (2006): AnnAGNPS – Annualized Agricultural Non-Point-Source Pollurant Loading Model.