VIRTUAL DATA SET – AN APPROACH FOR THE INTEGRATION OF INCOMPATIBLE DATA

Eva-Maria Stephan, Andrej Vckovski and Felix Bucher
Department of Geography
University of Zurich
Winterthurerstr. 190
CH-8057 Zurich, Switzerland

ABSTRACT

Data integration within GIS is made difficult by the incompatibility of source data. Here both, traditional approaches are discussed and an enhanced strategy for data integration is proposed. The concept of a virtual data set is presented, which allows more flexible and – due to quality information – more reliable integrated analysis. As a conclusion implementation issues are discussed, including the benefits of object-oriented techniques, comprehensive data quality modelling and visualization.

INTRODUCTION

This paper focuses on strategies for data integration. It presents a conceptual framework for data integration, which is an aspect of a multi-year program supported by the Swiss National Science Foundation, dealing with climate change, its impacts on ecosystems in alpine regions, and its perception by human beings. Investigating spatial variations of climate change and its impacts on ecosystems requires the use of models where information of various kind of GIS databases are interrelated. This paper first discusses the mutual relationships between GIS, data integration and environmental data modelling and analysis. It then shows current strategies for data integration with a subsequent evaluation of the inherent problems. Finally we present and discuss an alternate concepts for successful data integration.

DATA INTEGRATION AND GIS

In the past decades the use of computer-based means for processing of spatial information has significantly grown in importance. Particularly the development of Geographical Information Systems (GIS) has contributed to this evolution and, as a consequence, has expanded the potential for the analysis of spatial processes and patterns. At the same time data production has grown enormously, a phenomenon which is sometimes called the *data revolution*. Data is now drawn from various sources, gathered with different methods and produced by different organizations. For these reasons data are often not directly comparable in an integrated analysis with respect to their spatio-temporal references and attribute characteristics. The new facilities in spatial data processing on the one hand and the effects of data revolution on the other hand have accentuated the *data integration problem*. Data integration can be defined as the process *„of making different data sets compatible with each other, so that they can reasonably be displayed on the same map and so that their relationships can sensibly be analysed"* (Rhind et al., 1984).

GIS link together diverse types of information drawn from a variety of sources. Thus information can be derived *„to which no one had access before, and* [GIS] *places old information in a new context"* (Dangermond, 1989, p. 25). In fact, the ability of GIS to integrate diverse information is frequently cited as its major defining attribute and as its major source of power and flexibility in meeting user

needs (Maguire, 1991). Data integration facilitates more accurate analysis of spatial processes and patterns and encourages to use interdisciplinary thinking for geographical problem solving. Finally, data integration is the most important assumption for GIS to meet the expectations as a tool for decision support for planning tasks.

Data Integration in the Context of Environmental Information

In general, the investigation of natural phenomena is a highly interdisciplinary task. Particularly the interaction and dynamics of specific natural processes is not yet well understood and is of great interest in current research. Therefore, data integration is of special importance in the field of environmental data analysis. One can assume that the more interdisciplinary an analysis is, the more likely data integration will be a problem. Beyond that, many spatial phenomena are either difficult or expensive to measure or to observe, requiring their estimation from cheaper or more readily available sources. The substitution of such spatial phenomena, sometimes referred to as the derivation of secondary data, has become increasingly relevant as an application in the field of GIS. Examples are the calculation of the amount of soil erosion by means of the Universal Soil Loss Equation (USLE) or the simulation of the spatial distribution of vegetation patterns based on a model which takes topographic, climatic and edaphic factors into consideration.

Heterogeneity as a Bottleneck for Data Integration

Geographical entities are by definition described by their spatial, temporal and thematic dimension. At the point when different data sets enter an integrated analysis, one is often confronted with the problem that data have different characteristics according to these dimensions. To point out different characteristics between data sets the term *inconsistency*[1] has been established in the field of GIS.

Many of the problems of heterogeneity are a consequence of the fact that data are an abstraction of reality. Depending on the degree of abstraction and on the conceptual model for the transformation of reality into a *data set*, these characteristics can vary quite strongly. Beyond that, heterogeneity between data sets is also being introduced by different data gathering strategies, previous preprocessing steps and a lack of standardization.

Data Integration and Data Quality

The difficulties in data integration that accrue from heterogeneity are often reinforced by the uncertainty that is inherent to the data. Consequently, the process of data integration should strictly include data quality assessment for the resulting data set. The effects of uncertainty in individual maps on data integration is the subject of extensive research by Veregin (1989), Chrisman (1989), Maffini et al. (1989) and Lodwick (1989), among others.

In fact, environmental data are particularly affected by uncertainty. There are many reasons for that, including:

- Thematic surfaces of environmental characteristics may not be directly visible and therefore may not be verified at a reasonable expense.
- Many natural phenomena undergo continuous change (e.g. coastlines).
- Some natural phenomena cause problems because their realizations cannot be distinguished clearly and have transitional zones (e.g. vegetation).

---

1. Actually, the term 'inconsistency' can be quite confusing. Inconsistency has both the meaning of *not in agreement* and *to be contradictory*. In the present context, inconsistency refers exclusively to its first meaning. To avoid confusion, we relate here heterogeneity and heterogeneous data sets, respectively, to inconsistency.

- Due to the high costs of data gathering sample sizes of environmental data sets are normally much smaller than in other fields (e.g. terrain elevations for the generation of a Digital Terrain Model).
- Some of the environmental data cannot be measured directly in the field and have to be subsequently analysed with physical or chemical laboratory methods.
- Natural phenomena often have not negligible non-deterministic variability.
- Local deterministic variations as well as intrinsic variability of the data often cannot be exhaustively captured with the samples drawn.

OPERATIONAL DATA INTEGRATION

Traditional Approaches

Traditional approaches consider data integration as a two-step procedure (Shepherd, 1991): The first step tries to reconcile heterogeneous data sets to ensure their comparability. The second step involves the use of appropriate procedures to interrelate consistent data in a way that they meet the needs of a particular application. The reason for this two-step approach is that heterogeneity exists between the different data sets according to the intended application. In an *operational* system these two steps are separated from each other. In order to simplify future applications, the data sets are transformed into a pre-specified, *common format* when entering the system, such that comparability is ensured (see also figure 2). Ideally, the common format should be based on requirements of future analyses, however those can hardly be estimated exhaustively. Unfortunately, its specification usually is restricted by non-context-specific criteria like economic, hard- and software, and data limitations.

In the past few years, some broad strategies for specifying a common format have been adopted. In some applications, comparability is achieved by reducing diverse spatial or attribute data to some lowest common denominator representation. For example, all attribute data may be down-scaled to nominal information, or all spatial data may be converted into a grid of coarse resolution. Other applications may not accept the reduction of the variability in the source information and achieve comparability by transforming data sets according to the data set with the highest resolution. Other approaches include the conversion of all source data into a single target version, as in the integration of multiple data base schemata (Nyerges, 1989) or, the conversion to one single data model, as is the case of vector-only or raster-only GIS (Piwowar et al., 1990).

In any case one must be aware, that such transformations on the original data introduce further uncertainty.

Problems and Improvements

The previous section has given an overview of current approaches for solving problems of data integration. Regardless of the fact that these approaches are widely accepted and applied, we are of the opinion that they need further improvement.

Data integration usually does not include data quality assessment: The procedures for reconciling heterogeneous data sets are performed by means of traditional (mechanistic) transformations (e.g. interpolation, scale and projection transformations) and result in spatio-temporal references and attribute characteristics that are only apparently equivalent. These transformations all have in common that they involve prediction of attribute data at predefined locations and points in time. Because prediction generally introduces further uncertainty, data quality assessment and error propagation methods must be included to evaluate the reliability of the resulting data set and its limitations for a particular applica-

tion. Furthermore, data quality information of component data sets is a prerequisite for the estimation and monitoring of the effects of uncertainty on integrated data sets. Openshaw (1989, p. 263) notices that *„the effects of combining data characterised by different levels of error and uncertainty need to be identifiable in the final outputs"*.

Data should always be explained by means of meta information: Meta information improve the data integration process at various stages. However, a highly structured format is necessary for operational use. Burrough (1991, p. 172) notices that *„formalization of the knowledge that we already have and putting that in a knowledge base next to a GIS would help the user choose the best set of procedures and tools to solve the problem within the constraints of data, data quality, cost and accuracy."* Ideally, meta information should include declarations about: (a) the process of which the data set is a realization; (b) the conditions during data capture; (c) former applications and preprocessing of the data (the so-called lineage or history of data); (d) Characteristics of the present format; (e) Quality and reliability (based on the information of points a-c) and limitations for specific applications. Beyond that, meta information even should include specifications of functionality to avoid inappropriate user actions with the data.

The specification of the common format is rigid: To avoid future problems of heterogeneity, GIS systems and data bases are often designed so that they store all data sets in a common format to ensure their comparability. Presumably the (rigid) common format does not meet the requirements of all future applications. As a consequence, additional transformations of the data are likely to be needed, which will introduce further uncertainty. This is especially fatal when expert knowledge would be needed for the transformations or predictions applied, but such information is not stored with the data.

It is difficult or even impossible to avoid these problems, since the specification of the common format is restricted by non-context-specific limitations. Additionally, the full adequacy of a given common format for all future applications can hardly be achieved.

A MORE FLEXIBLE APPROACH TO DATA INTEGRATION

Definition of Heterogeneity and Prediction

In the previous section heterogeneity was introduced as a term describing the inconsistency (or incompatibility) of two or more data sets. This incompatibility needs to be defined accurately to point out its significance to the problem of data integration. We start out with a definition of the term heterogeneity with respect to data integration:

> *Data sets to be combined are called heterogeneous (with respect to the specific application) if some data values need to be predicted before performing the integrating operation.*

The operation thus requires data that are not present in the original data sets and implies the use of prediction methods to derive the missing information. These methods most often are needed to predict attribute values at spatial locations and/or points in time using the data available in the original data sets[1] (e.g. extra- or interpolation methods). This definition of heterogeneity also includes other types of predictors, which sometimes may be degenerate in the sense that predicted values are analytical functions of the input values. A simple example of such a degenerate predictor would be a method that predicts temperature values

1. Of course, it is possible, and often desirable, to include additional data sets to improve predictions. We refer to such data sets as *original data sets* as well.

in degrees of Fahrenheit given the temperature in centigrade or the transformation from one coordinate system to another. While usual predictors increase uncertainty in the predicted data, there are predictors which leave the uncertainty unchanged or even reduce it.

This extension of the term *prediction* to analytically deducible values allows the use of the above definition of heterogeneity in a broader sense when discussing integration problems.

The Idea of a Virtual Data Set

In this section we present an approach that can be used in an operational system to overcome heterogeneity and to better encapsulate the expert knowledge within the data set. This concept is termed *virtual data set*. It is a proposal for a more generic approach to data integration.

The basic idea of the virtual data set is the *extension of a data set with methods to provide any derivable or predictable information*. Instead of transforming original data to a standard format and storing them, the *original data are enhanced with persistent methods that only will be executed upon request*.

As the name indicates, a virtual data set contains virtual data. Virtual data is information that is not physically present. That is, it is not persistent[1]. This data is computed on request at run time.

As outlined before the traditional approach solves the problem of heterogeneity using a common format for the data. Instead of transforming the data to that common format the virtual data sets include methods for such transformations (which are predictions in our sense) together with the original (unchanged) data. Neglecting implementation and performance details, those two approaches are equivalent as long as the application of the prediction methods are transparent to the user. The second approach, however, can easily be enhanced to be more efficient. Once the transformation or prediction methods for getting the data into the common format are known, it is often easy to define methods that transform the data into yet another format. Suppose a transformation exists, that interpolates an original data set in a grid of 1 km resolution. It will not be very difficult to change this interpolation method so that it will produce data on a 0.9 km resolution grid instead. It might even be possible to parametrize the method enabling it to supply data in any parameter-dependent resolution. The step from specialized to more general predictors is often small. Anyway, a representation consisting of the original data together with prediction methods always contains equivalent or more information than the transformed data itself.

Once the application of the prediction methods is fully transparent[2] to the user, it is preferable to enhance an existing data set with prediction methods instead of transforming it using those methods. It is important to note, that these methods are designed to provide quality information for each predicted value. It would be even favourable to have the quality information being an inherent part of both, virtual and original values.

Figure 1 shows a schematic view of an original data set, its transformation according to a common format and its enhancement to a virtual data set. The data set shows the spatial distribution of a certain phenomenon (attribute A) at given locations (s_1, s_2, s_3). The analysis requires an additional attribute B which can be derived from A. The common format specifies the locations s_a, ..., s_f where the attributes are interpolated. The virtual data set is equipped with methods to interpolate A at any location, to derive B from A and to transform spatial references between different coordinate systems.

1. We refer to persistent data if they are available on secondary (external) storage.
2. It should make no difference whether the user accesses data really available in the data set or virtual data that has to be computed using the prediction methods first.

Figure 1: Static structure diagram of the homogenization of a data set using a common format and a virtual data set.

Comparing the virtual data set with traditional approaches one can see that it provides more flexibility, since there are no constraints by a given common format. Beyond that, a virtual data set is designed such, that data quality information is mandatory. The virtual data set enables the user to carry out any integrated analysis and see the effects of missing or unpredictable data as uncertainties visualized on the resultant maps; whereas traditional approaches limit the integrated analysis to data available in a common format, as long as the user does not perform additional transformations. The idea of a more flexible approach is also encouraged by the complexity of the operations involved particularly in environmental decision support. It is often very difficult or even impossible to estimate the influence of different input values to the result without the help of error models or sophisticated sensitivity analysis. Even very uncertain predictions may be of more value in an integrated analysis than no value at all.

The virtual data set also encapsulates the expert's knowledge (choice of the appropriate methods, meta information) within the data set, which will presumably lead to better predictions and reliable quality informations.

The virtual data set is able to provide values at very high resolutions when requested. So it is important to avoid the common misunderstanding that data of high resolution are implicitly more accurate than data of coarse resolution.

Figure 2 compares the data integration process with respect to the data flows in the traditional approach and the virtual data set for a sample integrated analysis.

SOME REQUIREMENTS FOR USING VIRTUAL DATA SETS IN GIS

Having introduced the general idea and some theoretical background of data integration on the basis of virtual data sets, we will now concentrate on some implementation issues. Commercial GIS do not meet the requirements for handling virtual data sets. An essential need is comprehensive data quality handling including error propagation and data quality visualization. Furthermore, the use of an object model will simplify the implementation of the virtual data set concept.

Object Model and Persistence

One of the key ideas of the virtual data set is the need to have the prediction methods stored with the original data. This, together with the needs for high level information hiding or encapsulation, suggests the application of the object model and object-oriented design (Booch, 1991) for describing the structure of the virtual data set.

A general problem is the required persistence (i.e. storage) of the prediction methods. One of the motivations of the virtual data set was the encapsulation of expert knowledge (e.g. prediction methods) within the data set. This will enable an application-independent use of the data. Data and procedures must thus be included in a data set. While current data base management systems (DBMS) offer little support for procedural data, especially when they have to be stored together with the 'normal' data within the data base, there are some approaches to enhance an (object-oriented) DBMS to allow storage of procedural data (e.g. Deux, 1991; Heuer, 1992). It is, however, difficult to establish similar capabilities for persistence and transfer between systems for procedural data since procedural data often depend heavily on characteristics of processors, compilers or interpreters. Some promising work adressing the integration of distributed systems with an *object request broker architecture* is presented by the Object Management Group (Soley, 1992).

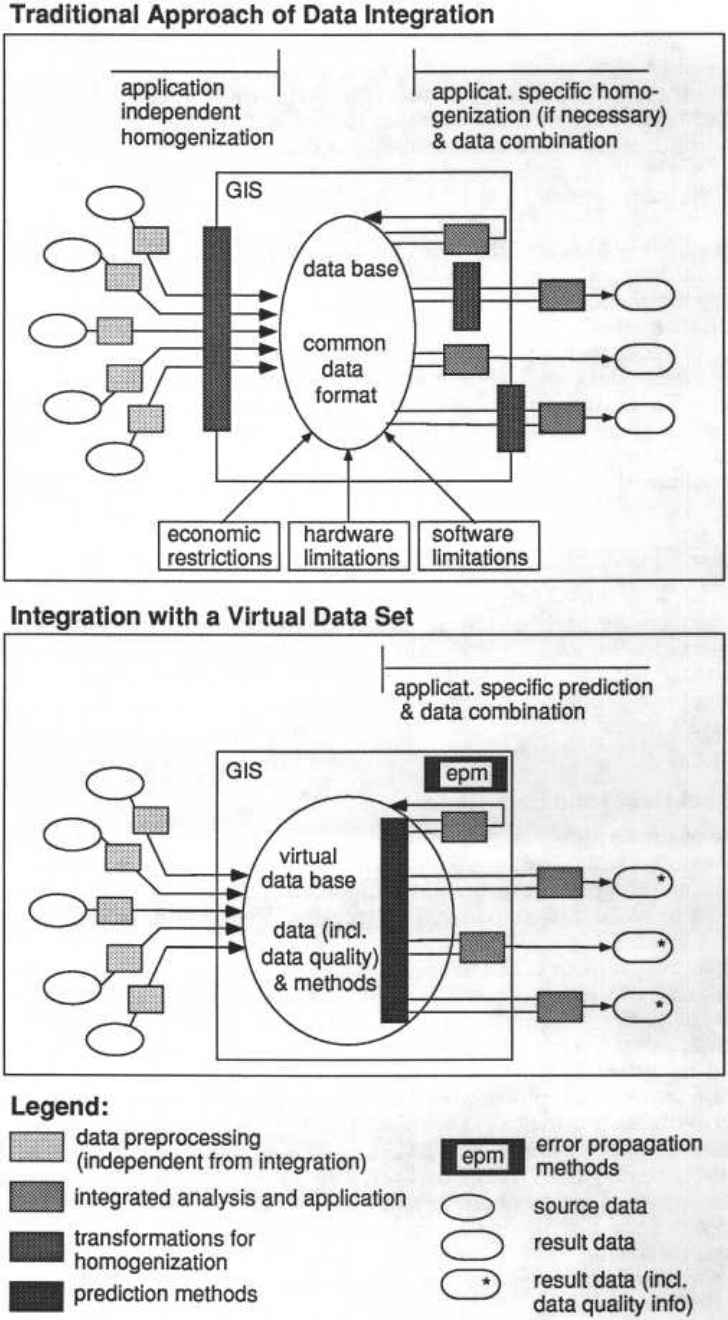

Figure 2: Comparison of data integration process in traditional approach (upper) and virtual data set (lower).

Error and Uncertainty Propagation

In the previous sections we always assumed that GIS are capable of handling data quality information transparently. This assumption is not very realistic. The heterogeneity of the data sets and the complexity of the operations performed demand more efficient and highly integrated capabilities for assessing uncertainty.

The concept of virtual data sets increases the importance of uncertainty handling by adding new sources of uncertainty. While many operations are forbidden in a traditional integration approach because of non-existent data, the virtual model allows virtually any operation between two or more data sets. The difference is, that the virtual data set might deliver absolutely uncertain values at locations or points in time when there is no reliable prediction possible.

Currently, considerable research efforts are on the way with respect to data quality and error propagation models in operational GIS. For example, Wesseling and Heuvelink (1991; 1993) present a system based on second order Taylor series and Monte Carlo methods among others to estimate error as a result of operations on stochastic independent and dependent uncertain input values.

In addition to those ideas we suggest the use of interval mathematics (e.g. Moore, 1979) for an easy to implement and very conservative error propagation scheme (i.e. error is never under-estimated, but often over-estimated). Especially for environmental applications the complexity of the problems encourages the use of conservative error estimates.

Interactive Visual Support for Data Integration

Interactive visual support is an important component of data integration. In order to be able to use GIS as a decision support system, it should be highly interactive and present graphical results. On the one hand the system needs to acquire expert knowledge in a communicative and exploratory way, such as the decisions leading to the selection of appropriate prediction methods to create a virtual data set.

On the other hand the system should be capable of visualizing data quality. While the need is clear from the above considerations, the solution is not. Data quality information adds several new layers of information in a GIS data base. Its visualization needs, therefore, to be able to handle multidimensional data.

Verification and validation of the selected or newly defined prediction methods should be supported in a standardized way. The subsequent users of the (virtual) data set will usually not reflect on the prediction methods defined and are therefore dependent on a consistent quality of the prediction methods[1].

CONCLUSIONS AND FUTURE RESEARCH

We have shown that major problems of data integration are heterogeneity and the lack of comprehensive data quality handling. The homogenization of data sets always involves prediction and thus adds further uncertainty. A flexible homogenization scheme is established with the help of the presented virtual data set. Enhancing it with the appropriate data quality models will facilitate an unrestricted, yet reliable integrated analysis.

Future research will concentrate on refinement of this concept and its realization. This demands a detailed application of the object model and the embedding of prediction methods and error propagation models.

---

1. It is very important to note that the *consistent quality* does not stand for the quality of the predicted virtual data. Rather, it guarantees that the prediction *method* meets a certain minimum quality requirement. This provides a kind of quality assurance during the process of the defining the virtual data set.

ACKNOWLEDGEMENTS

This work has been funded by the Swiss National Science Foundation under contract No. 50-35036.92. We would also like to acknowledge the contributions of Prof. Dr. Kurt Brassel and Dr. Robert Weibel.

REFERENCES

Booch, G. 1991, *Object oriented design with applications*, Benjamin/Cummings Publishing Company, Redwood City

Burrough, P.A. 1991, The Development of Intelligent Geographical Information Systems: *Proceedings of the EGIS'91 Conference*, Brussels (Belgium), pp. 165-174.

Chrisman, N.R. 1989, Modeling error in overlaid categorical maps: Goodchild, M.F. and Gopal, S., *Accuracy of spatial databases*, Taylor & Francis, London, pp. 21-34.

Dangermond, J. 1989, The organizational impact of GIS technology: *ARC News*

Deux, O. 1991, The O_2 system: *Communications of the ACM*, Vol. 34, pp. 34-48.

Heuer, A. 1992, *Objektorientierte Datenbanken: Konzepte, Modelle, Systeme*, Addison-Wesley, München

Lodwick, W.A. 1989, Developing confidence limits on errors of suitability analyses in geographical information systems: Goodchild, M.F. and Gopal, S., *Accuracy of spatial databases*, Taylor & Francis, London, pp. 69-78.

Maffini, G., Arno, M. and Bitterlich, W. 1989, Observations and comments on the generation and treatment of error in digital GIS data: Goodchild, M.F. and Gopal, S., *Accuracy of spatial databases*, Taylor & Francis, London, pp. 55-68.

Maguire, D.J. 1991, An overview and definition of GIS: Maguire, D.J., Goodchild, Michael F. and Rhind, David W., *Geographical Information Systems: principles and applications*, Longman, London, Vol. 1, pp. 9-20.

Moore, R.E. 1979, Methods and applications of interval analysis: Society for industrial and applied mathematics, *Studies in applied mathematics*, Vol. 2

Nyerges, T.L. 1989, Schema integration analysis for the development of GIS databases: *Int. Journal of Geographical Information Systems*, Vol. 3, pp. 153-183.

Openshaw, S. 1989, Learning to live with errors in spatial databases: Goodchild, M.F. and Gopal, S., *Accuracy of spatial databases*, Taylor & Francis, London, pp. 263-276.

Piwowar, J.M., Le Drew, E.F. and Dudycha, D.J. 1990, Integration of spatial data in vector and raster formats in a geographic information system environment: *Int. Journal of Geographical Information Systems*, Vol. 4, pp. 429-444.

Rhind, D.W., Green, N.P., Mounsey, H.M. and Wiggins, J.S. 1984, The integration of geographical data: *Proceedings of the Austra Carto Perth Conference*, Perth, pp. 273-293.

Shepherd, I.D.H. 1991, Information integration and GIS: Maguire, D.J., Goodchild, Michael F. and Rhind, David W., *Geographical Information Systems: principles and applications*, Longman, London, Vol. 1, pp. 337-360.

Soley, R.M. 1992, Using object technology to integrate distributed applications: *Proceedings of the First Intern. Conference on Enterprise Integration Modelling*, Hilton Head, SC (USA), pp. 445-454.

Veregin, H. 1989, Error modeling for the map overlay operation: Goodchild, M.F. and Gopal, S., *Accuracy of spatial databases*, Taylor & Francis, London, pp. 3-18.

Wesseling, C.G. and Heuvelink, G.B.M. 1991, Semi-automatic evaluation of error propagation in GIS operations: *Proceedings of the EGIS'91 Conference*, Brussels (Belgium), pp. 1228-1237.

Wesseling, C.G. and Heuvelink, G.B.M. 1993, Manipulating quantitative attribute accuracy in vector GIS: *Proceedings of the EGIS'93 Conference*, Genoa (Italy), pp. 675-684.