MULTISOURCE, MULTINATIONAL ENVIRONMENTAL GIS: LESSONS LEARNT FROM CORINE

H M MOUNSEY

This chapter complements those by Clark, Hastings and Kineman and by Townshend in discussing the design and creation of environmental databases covering large areas. It differs, however, in describing the problems associated with databases created largely from existing data sets, many derived from map and on-ground sample sources. The problems of defining user needs in multinational projects and the consequent difficulties of system design – best approached by prototyping – are outlined. Throughout, the arguments are illustrated by examples drawn from the European Commission's CORINE programme: this multinational environmental monitoring and assessment tool was set up from 1985 onwards and was predicated entirely upon the availability of a comprehensive database and GIS for the 12 countries in the Community. Experience gained in this programme has fostered moves to harmonize the collection of much environmental data and hence minimize the variations in parameters encountered across national frontiers because of differences in the collection methodology. A key factor in the programme was the requirement to meet the changing needs of the bureaucracy in Brussels: the paramount need for system designers, environmentalists and bureaucrats to understand each other is stressed. The project's success has led to agreement to set up a European Environment Agency.

INTRODUCTION

Rising concern over the degradation of the environment has resulted in an increase in research on the identification and study of environmental problems. Unfortunately, much of this work has been speculative and theoretical and, at least until recently, not supported to any great extent by adequate databases. This situation is changing: in parallel to a rapid rise in the volumes and quantity of data collected, massive changes in technical capability have facilitated the development of GIS to handle the diversity of information involved. Moreover, since environmental planning and management is inherently cross-disciplinary, the use of GIS technology to build environmental databases from disparate sources of information to study problems of some commonality is highly appropriate.

The forerunner of much work in the development of national environmental databases was the now well-known Canada Land Inventory (Canada Department of Forestry and Rural Development 1965; Tomlinson, Calkins and Marble 1976), since followed by the developmental work of the US Environmental Protection Agency (EPA 1987) and the proposed development of an Australian federal resources database (Mott 1990). Environmental problems, however, are not only a matter for national concern; they also have

profound social and economic consequences at a continental and global scale. Examples of this include the widespread effects of environmental disasters like the Chernobyl explosion, famine throughout the African sub-continent and the late-1980s drought in the American Midwest. Further, although their effects are not as well understood, processes operating at a global scale (e.g. those leading to the greenhouse effect and the ozone hole) are now also recognized as having significant local impacts; the development of environmental databases which allow further study of such complex real- and whole-world problems is now both possible and necessary.

Environmental databases are being developed by a number of organizations to address a wide range of issues (Clark, Hastings and Kineman 1991 in this volume; Townshend 1991 in this volume); there is much diversity in scale and spatial coverage, in technological implementation, in the range of data holdings and in the organizational background supporting the development and use of the databases. However, some features common to all operational environmental databases covering large areas can be identified:

- They typically draw on a wide range of spatial data sources (i.e. data with some type of locational reference).
- They provide software for data retrieval, modelling and output by a wide range of users of varying abilities.
- They normally operate centrally within a corporate organization – generally as a spin-off from other activities (indeed, it is often doubtful whether it would be economic to develop such a database solely for environmental monitoring purposes).

What then are the major issues behind, and the challenges facing, the development of environmental databases? One way to examine these is within the context of an existing continental-scale environmental database, the CORINE (CO-ordinated INformation on the European environment) Programme. The development of this programme has been supported by the European Community (the EC) since 1984 and, in nearing the conclusion of its development phase, it provides a good example of the development of a database from initial idea to working prototype (CEC 1990). Many of the issues which had to be resolved over the early period of the Programme are now well understood and even pedestrian; moreover, developments in hardware have transformed some problems from daunting to trivial. None the less, the early stages of CORINE remain a good example of GIS implementation because of four factors: the continuing need for pragmatism in building databases near the limits of contemporary technology; the commonplace need to sew together data from many sources; the essential requirement to provide a database useful to the bureaucracy; and – at the same time – the desirability of ensuring that the host of pragmatic decisions did not render the results of analyses meaningless in scientific terms. Finally, by way of introduction, this chapter differs from those by Clark *et al.* (1991 in this volume) and Townshend (1991 in this volume) because they concentrate on the use of global or continental environmental databases assembled largely from remote sensing imagery; relatively little of CORINE data has been obtained from such sources to date. Despite this, the reader is urged to read all three chapters to obtain a comprehensive picture of environmental GIS applications.

DEVELOPMENT OF THE DATABASE

The establishment of a database to meet the requirements of a user community normally follows a well-defined series of steps irrespective of the subject matter of the database (Tschritzis and Lochovsky 1982; Benyon 1990). Simplified, these include:

- identification and documentation of the user requirements;
- definition of the data requirements which will address the user requirements;
- establishment of an information technology (IT) solution which offers facilities for data handling to meet the user requirements;
- an assessment of the costs and benefits of such a solution; and

- if the above are favourable, installation and implementation of the selected solution.

In building databases to meet business requirements in the government, commercial and utilities sectors, such steps are usually reasonably easy to define and follow. However, in the establishment of environmental databases (and in this the CORINE Programme is no exception), the path is not as clear for a number of reasons. For example, it is often difficult to delimit the range and number of user requirements and to rank them in order of importance. The CORINE Programme in particular is broad in scope and its outer boundaries hard to define; ultimately, the database will serve a much wider range of applications than those defined at the outset. While these original foci may serve as a starting point from which to define data requirements, it seems likely that – as in the case of many environmental databases – it is the availability of existing data which (initially at least) dictates the range of applications rather than *vice versa*; the first two stages above are thus effectively reversed. Nevertheless, it is convenient to discuss each stage in turn and this is done below.

Definition of user requirements

The origins of the CORINE Programme lie in studies during the late 1970s towards the establishment of an environmental database for the European Community (Rhind *et al.* 1986). The Programme itself was formally established by the Directorate General of the Environment (DG XI) in June 1985 and was aimed at 'gathering, coordinating and ensuring the consistency of information on the state of the environment and natural resources' as an aid to Community environmental policy (Official Journal of the European Community 1985). Such aims are rather broad in scope; thus two more specific tasks were targeted for study. The first of these was seen as the improvement of data availability and compatibility both within the European Community itself and in the member states, to be achieved through the development of appropriate techniques for the collection, storage, manipulation and output of environmental data. Secondly, in order to focus data collection policies and to avoid the random assimilation of data, three specific topics of environmental importance were identified as initial targets for study. These include:

- biotopes (through the setting up of an inventory of sites of scientific importance for nature conservation);
- acid deposition (through provision of information on emissions and on risks of damage to flora and fauna, etc.);
- protection of the environment in the Mediterranean region (through the supply of information on land cover, quality and use, on water resources and on coastal problems).

Several of working groups of national experts were established to address these and related topics; these groups included those on air pollution, on coastal erosion, on water resources, on land use in the Mediterranean region and on biotopes. Each group defined the nature of the problems to be solved and the data requirements to address them, organized the acquisition of data from pre-existing sources and fed the data sets to a centralized database for analysis and output to other users. The user community for these data was seen in the first instance as being limited to DG XI but with subsequent expansion to other DGs within the European Commission, to international organizations such as the UN Environment Programme and the UN's Food and Agriculture Organization, and eventually to *bona fide* users within the member states such as government institutes and individual researchers.

Definition of data requirements

The range of data to be included in an environmental database depends naturally on the objectives of the system. In the case of the CORINE database, the three specific topics of environmental importance listed in the original communiqué (see above) should have defined, at least in part, the range of data for inclusion. However, these were broad in scope and left much room for interpretation. Thus, in practice, data collection has been governed by pragmatic considerations: the constraints of resources and time, of data availability and consistency, and of data volumes. These practical limitations should not

be underestimated, especially when designing an environmental database to meet the many and diverse requirements of the European Community; its land area is 2.25 million square kilometres and attempts to cover this in fine detail would have generated quantities of data which were unmanageable – at least when the programme began. In the development system at least, such detailed data would have been both unwieldy to use and costly to administer. Furthermore, time and resource availability precluded the primary collection of data by ground survey to any great extent. Thus, at the outset of the Programme, four basic principles were defined for data collection (Wiggins *et al.* 1987):

- that raw data (as opposed to aggregated or interpreted data) should be included as far as possible, allowing for maximum use by researchers wishing to carry out their own classifications and aggregations to meet their specific needs;
- that existing data should be used wherever possible;
- that data input from maps should be restricted to small scales (1 : 250 000 or less) in order to reduce data volumes to manageable levels, at least during the early stages of the Programme, and to minimize international compatibility issues and to keep the data conversion costs within acceptable bounds;
- that only data already in machine-readable format should be used as far as possible, to minimize the need for encoding and digitizing.

Clearly these constraints reduced the amount of data available and are not wholly attainable; in order to obtain any data whatsoever on some topics, some encoding and digitizing had to be undertaken and undesirably aggregated data had to be included in the database. It is also unrealistic to study other environmental issues (for instance, coastal erosion) at such low resolution; hence, in practice, data collection at map scales as large as 1 : 25 000 has been undertaken – in rare circumstances and with consequential difficulties, as discussed below.

It is inappropriate to consider the holdings of the CORINE database in detail, not least because they are constantly under review. However, Table 48.1 provides an overall view of the holdings and Whimbrel (1989) and CEC (1990) document the holdings in greater detail. A fundamental requirement of any environmental database is a sound topographic framework, to which all other data can be related; in effect, it acts as a spatial template or control mechanism ensuring spatial consistency. Ideally, this should include, as a minimum, the coastal outline for display purposes. For environmental modelling, however, there is normally an additional requirement for information pertaining to the hydrological network, to ground altitude and to slope. Unfortunately, the CORINE Programme encountered significant difficulties in obtaining such data for the European Community. These did not arise from a simple lack of information; indeed, some digital topographic data exist for most of the member states of the Community at national level. But the characteristics of the data sources from which these were derived (the map scales and projections), the data contents (which features have been included, the contour intervals used, etc.) and the degree of topological structuring all differ so markedly that integration into a common, small-scale database, at least in the short term, would have been an impossible task. Furthermore, there is no common official map series across the European Community at a scale greater than 1 : 1 million (and, for areas of Greece, no topographical maps are available at all for reasons of military confidentiality); thus the digitizing of a topographic base would have been a considerable, if not an impossible, undertaking.

The solution adopted is, in many ways, less than ideal but at least it has provided a topographic base for the CORINE Programme. A 1 : 1 million scale digital database for the European Community, but excluding Greece and originally digitized from the ONC (Operational Navigation Charts) series, was obtained from the German national mapping agency, the Institut fur Angewandte Geodasie (IfAG). Received in 'spaghetti' form, this was topologically structured at Birkbeck College, University of London and had several hundred digitizing errors removed; with the addition of data for Greece (digitized in-house from the ONC series), this forms the initial CORINE topographical framework. Primarily designed for use in air navigation, the topographic base of the ONC series is not of the highest quality (see Rhind and Clarke 1988 for some examples of its internal

inconsistencies) but nevertheless it formed the best single, consistent data set available. A number of additional layers of thematic data have been added to this base; these include a digital representation of the European Community's soils map of Europe compiled at 1 : 1 million scale, climatic data compiled from national meteorological organizations and information on biotopes, coastal erosion, land cover and water resources, compiled from local sources by groups of national experts.

The IT solution

The wide variety of tasks to be addressed within the CORINE Programme and the potentially enormous volumes of data suggested at the outset the need for a powerful GIS. The requirements for other software and for hardware were less well defined although some were identifiable in outline; these included the requirement to:

- be easy to use and maintain;
- be a relatively low-cost solution;
- be flexible enough to handle large volumes of data from many sources, at a wide range of scales and in many projections; and
- offer full GIS functionality to perform the many routine tasks required of an environmental database.

The selection of any IT system, including a GIS, would usually be subject to full testing and bench-marking procedures in order to establish the optimum from various possible alternatives. However, in order to launch the experimental programme in a short time period and to permit evaluation of different options, it was thought appropriate to use an existing system and expertise. Thus a pilot system was established, based on ARC/INFO at Birkbeck College in the University of London, UK; although not necessarily intended as a long-term solution, this prototyping has indeed provided invaluable pointers towards the long-term requirements as well as supporting the short-term needs. The same software was later implemented in DG XI in Brussels.

Functional requirements of a GIS

Data input

Data were generally input to the CORINE database from two sources: third-party data already in digital format or through digitizing maps by EC staff or a sub-contractor. In either case, data capture – and subsequent validation and editing – is time consuming but generally provides little technical challenge. On the other hand, data input frequently includes the process of data conversion, notably projection conversion and generalization. These are necessary because source maps often vary in their scale and projection, while data from existing databases are often provided in a wide range of geographical forms; unless data are converted onto a consistent spatial base, accurate data integration is not possible (Flowerdew 1991 in this volume). As well as algebraically-based projection facilities, provision for the 'rubber sheeting' of input data has proved essential; in a number of cases, the projection of the input map or data was unclear, unknown or even specified incorrectly; in such circumstances, local transformations were applied to give a 'best fit' to other data sets using large numbers of control points as a spatial template. It follows that detailed documentation of the procedures applied to each data set was essential.

Some databases are considered by the end user to be 'scale free' (in that they can be output at any scale within the constraints of the user's hardware). In practice, however, the storage of 'scale free' databases is still at a research stage (Muller 1991 in this volume). Thus within the CORINE Programme, data are stored in one of two forms appropriate to the scales most usually required by the end user, one at a notional scale of 1 : 1 million and another at 1 : 3 million. Moreover, in order to avoid massive storage volumes and long processing times, generalization procedures form an important feature of the exploitation of GIS software both in pre-archiving processing and at run-time.

Data analysis

It is self-evident that, if the data stored within an environmental database are to be of any use, the software must offer a capability to analyse them according to the user's requirements; the degree of matching between the analytical requirements of the user and the facilities offered by the system is often the most important criterion in the selection

Table 48.1 Overview and contents of the CORINE GIS.

Theme	Nature of information	Characteristics of digital data	Mbyte	Resolution/scale
Biotopes	Location and description of biotopes of major importance for nature conservation in the Community	5600 biotopes described on about 20 characteristics. Boundaries of 440 biotopes in Belgium and Portugal	20.0 2.0	Location of the centre of the site
Designated areas	Location and description of areas classified under various types of protection	13 000 areas with 11 attributes. Computerised boundaries of areas designated in compliance with article 4 of EEC/409/79 directive on the conservation of wild birds	6.5	Location of the centre of the site 1/100 000
Emissions into the air	Tons of pollutants (SO2, NOx, VOC emitted in 1985 per source category: power stations, industry, transport, nature, oil refineries, combustion	1 value per pollutant, per category of source and per region, plus data for 1400 point sources i.e. +/−200 000 values in total	2.5	Regional (NUTS III) and location of large emission sources
Water resources	Location of gauging station. Drainage basin area, mean and minimum discharge 1970–85 for southern part of EC	Data recorded for 1061 gauging stations for 12 variables	3.2	Location of gauging stations
Coastal erosion	Morpho-sedimentological characteristics (4 categories), presence of constructions, coastal evolution characteristics, erosion, accretion, stability	17 500 coastal segments described	25.0	Base file 1/100 000 generalized version 1/1 million
Soil erosion risk	Assessment of potential and actual soil erosion risk by combining 4 sets of factors: soil, climate, slopes, vegetation	180 000 homogeneous areas (southern part of Community)	4000.0	1/1 million
Important land resources	Assessment of land quality by combining 4 sets of factors: soil, climate, slopes, land improvements	170 000 homogeneous (southern part of Community)	300.0	1/1 million
Natural potential vegetation	Mapping of 140 classes of potential vegetation	2288 homogeneous areas	2.0	1/3 million
Land cover	Inventory of biophysical land cover in 44 classes	Vectorized database for Portugal, Luxembourg	51.0	1/100 000
Water pattern	Navigability, categories (river, canals, lake, reservoirs)	49 141 digitized river segments	13.8 0.3	1/1 million 1/3 million
Bathing water quality	Annual values for up to 18 parameters, 113 stations for 1976–86, supplied in compliance with EEC/76/160 directive	2650 values	0.2	Location of station
Soil types	320 soil classes mapped	15 498 homogeneous areas	9.8	1/1 million

Table 48.1 *Continued*

Theme	Nature of information	Characteristics of digital data	Mbyte	Resolution/scale
Climate	Precipitation and temperature (+incomplete data for other variables)	Mean monthly values for 4773 stations	7.4	Location of station
Slopes	Mean slopes per square km (southern regions of Community)	1 value per km^2 i.e. 800 000 values	150.0	1/100 000
Administrative units	EC NUTS(Nomenclature of Territorial Units for Statistics); 4 hierarchial levels	470 NUTS digitized	0.7	1/3 million
Coasts and countries	Coastline and national boundaries (Community and adjacent territories)	62 734 km	0.3 3.2	1/3 million 1/1 million
Coasts and countries	Coastline and boundaries (planet)	196 countries	1.5	1/25 million
ERDF regions	Eligibility for the Structural Funds	309 regions classified	0.01	Eligible regions
Settlements	Name, location, population of urban centres >20 000 people	1542 centres	0.1	Location of centre
Socio-economic data	Statistical series extracted from the SOEG-REGIO database	Population, transport, agriculture, etc	40.0	Statistical Units NUTS III
Air traffic	Name, location of airports, type and volume of traffic (1985–87)	254 airports	0.1	Location of airport
Nuclear power stations	Capacity, type of reactor, energy production	97 stations, up-date 1985	0.03	Location of station

(Source: CEC 1990).

of a GIS (Clarke 1991 in this volume). Only its use ultimately justifies the development of the system! Although the requirements of the end-user of the CORINE Programme were initially ill-defined, it is nevertheless possible to identify some basic requirements of such a GIS. These include facilities for feature selection and display, and for statistical analysis and modelling of single and multiple data sets (see Maguire and Dangermond, 1991 in this volume for further discussion of the functionality of GIS).

Feature selection and display includes selection both by geographical area and by thematic attribute. Where appropriate, this may also include generalization for mapping at smaller scales, including the generalization of attributes (e.g. the merging of classes or of individual features with specific attributes for clarity) according to predetermined rules. The overlay of separate data sets to produce a single data set with a combination of attributes is often important in environmental modelling, as is the construction of 'buffer zones' or 'corridors' of user-selected width around features of a defined type within data sets.

Data output

The results of any analysis must also be available in a form selected by the user. Typically, these might include tables and tabular reports but also a wide variety of graphics, produced either on a terminal or as hard copy. The production of a well-designed and balanced map is a much neglected area of GIS, given the importance of these in the communication of the results of an analysis to the end-user and the

problems arising from a lack of cartographic skills among many GIS users (Blatchford and Rhind 1989). The situation is a delicate one; on the one hand, a poorly designed map may fail to convey the results to the user and may also inadequately represent the effort involved in the establishment of the database and carrying out of the analysis. On the other hand, it is often easy to convince end-users on the basis of inadequate evidence – a highly effective map may well be used to mask inadequacies in the original data from the decision maker. Facilities must of course be available within the GIS to enable the production of well-designed maps, but the responsibility ultimately rests with both the producer and the user to ensure proper interpretation; there is, then, a moral and professional element to the use of GIS.

Costs versus benefits

The balance of costs and benefits is an extremely difficult one to establish for geographical databases of all kinds (Didier 1990; Calkins 1991 in this volume; Clarke 1991 in this volume). It is especially so for environmental databases (and was never undertaken formally for the CORINE Programme). The costs of the IT hardware and software will be the easiest to establish and data costs will be governed by data availability and hence are (usually) quantifiable. However, a major component in the costs of any programme are those of staff, the requirements for which are governed, at least in part, by the volume of use of the database and by the skills of the end-user. In addition, training and documentation needs often form a significant proportion of the total costs.

Benefits are even harder to quantify; the usual ones of improvements in service and productivity or exploitation of new business opportunities may be inappropriate measures in the creation of multinational databases. In such circumstances, the usual criteria are replaced by more intangible concepts such as 'better management of information and assessment of risk'. The implementation of a commercial strategy (and, therefore, the acceptance of the burden of cost) by one organization for the benefit of a wider user community requires that well defined cost recovery procedures are agreed beforehand. If not (as in the case of the CORINE Programme), the establishment of an environmental database is likely to be an act of faith investment legitimated for the greater good of the world's population, with largely intangible (or at least unquantifiable) benefits.

IMPLEMENTATION — THE REALITIES

Detailed planning of the construction of an environmental database is an idea which is excellent in theory (and in hindsight), but in reality is unrealistic; it is difficult to gauge the full measure of the user requirements and thus the data requirements that underpin these. In such circumstances, investment appraisal becomes a matter of academic speculation. The CORINE Programme has grown thus far through the enthusiasm of a small group of people and through the availability of appropriate technology, rather than through a well-thought-out development plan to meet the end-user requirements; it has also benefited from external shifts in policy and in public opinion. It is not unique in this approach.

The lack of a development plan aside, it is still possible to draw a number of lessons from the implementation of the Programme thus far; principally that the constraints on the development of environmental databases are not at present technology based, but relate to the availability of data and to aspects of access and use of the database. Areas of technical development on the research agenda for environmental databases include:

- the development of scale-free databases such that environmental issues should be addressed at local, regional, national, continental and global scale;
- the efficient and well-integrated handling of raster and vector information to ensure best use of all sources of environmental data;
- the recognition and handling of error conditions in analysis and modelling, especially in the light of the 'fuzzy' nature of much environmental data; and
- the development of icon-based interfaces to GIS

to enable the wider use by an increasingly non-specialist audience.

Data limitations

Although few system limitations have been encountered in the development of the CORINE database, the same cannot be said of the data. In a 'perfect' database, all layers of data would be spatially and temporally complete and consistent in terms of units of measurement, definitional, spatial and temporal characteristics (Briggs and Mounsey 1989). Even though the CORINE Programme is still under development, it is possible to highlight a number of problems which are representative of environmental databases in general. These include data availability and access, data quality, data maintenance and update, data volumes and data documentation.

Data availability and access

There are still many deficiencies in the CORINE database, both regional (e.g. the lack of adequate topographical data for Greece) or thematic (e.g. only limited data are available on atmospheric emissions for the whole Community). In addition (and not surprisingly in a database designed only to cover part of a continent), there are substantial edge-effects where data end at national borders. As an example of the latter problem, the lack of data for areas outside, but adjacent to, the EC prevented much work on either the Chernobyl explosion or on the consequences of a Swiss toxic spill into the Rhine. Both these and internal gaps in the data are a serious constraint which may take much time and resources before they can be overcome; experience in CORINE suggests that the '80 : 20 rule' may well apply (i.e. the last 20 per cent of the data required costs 80 per cent of the total effort). Even if the effort is discounted, the extension of a data set to cover adjoining countries represents a major policy decision and may have political ramifications.

Notwithstanding such deficiencies, the progressive 'bottom-up' development of an environmental database has the advantage that it may be a sensitive indicator of which data are already available and what else is required; if the data sets are available, but are not to be integrated in the database, it is still advantageous simply to know of their existence. Thus one valuable product of the CORINE Programme is an ongoing inventory of environmental data sources (CEC 1990), which is of use in its own right. An extreme example of such data 'signposting' or cataloguing (Department of the Environment 1987) is the Australian National Resources Information Centre, which is presently under development: it aims to hold no data at all, merely acting as a source of information on data holdings at state and federal level (Mott 1990).

Acquisition of data sets that are available in digital form is not always straightforward. In common with many other environmental databases, the CORINE Programme has never had a large budget with which to purchase data. Consequently, some available data sets were simply too expensive to be funded by viring from other funds. A related problem is that of transfer formats; the CORINE database draws upon data from a wide range of sources and thus international developments in data transfer formats are of particular concern to it. However, notwithstanding the existence of a number of national standards, it is difficult to force 'data donors', who may be contributing data on a very low or no-cost basis, to reformat data from their own method of organization to that requested by the builders of another database. Thus, notwithstanding the evolution of various standards (see, for instance, Guptill 1991 in this volume), at least in the foreseeable future staff involved in building any multi-contributor database will need to be adept at writing short 'one off' programs to reformat data from the many and varied formats in which they are received.

Data quality

By far the greatest problem in the development of any environmental database is that of data quality – are the data an accurate representation of the real world? Because the CORINE Programme, in common with many other environmental databases, draws on a wide variety of sources, the potential for variation in data quality and character is great. Variations in timeliness, spatial coverage, density and measurement method may all be hidden behind imprecise definitions and inconsistent use of terminology; alone or in combination, these all present a real danger to the end-user. More seriously, but even less considered, it is unclear how much liability rests with which party should unfortunate consequences arise from the use of such

data (Epstein 1991 in this volume). It is only through the understanding of the totality of these issues that the user can judge whether the data are appropriate and should be applied to his or her task.

There are two major components of data quality – accuracy and completeness (Chrisman 1991 in this volume). Within any one data set, these components must be known for both the position of the features and for their attributes. Positional accuracy is a measure of the proximity of the coordinates of any feature in the database to their true position on the ground and is a function, at least in part, of the method for data collection. Many of the data within the interim CORINE database are derived from digitizing paper maps; while positional accuracy is partly a reflection of the quality of data compilation, it is thus primarily dependent on the original map scale and the quality of the generalization carried out by a cartographer. Maps are usually held to be accurate to one line width (typically drawn at about 0.5 mm: Fisher 1991 in this volume); as source material used within the interim database was compiled within the range 1 : 500 000 to 1 : 1 million scales, this is equivalent to a maximum positional error of 250 m. However, while national standards laying down acceptable levels of generalization and accuracy frequently exist for topographical map compilation at medium scales (e.g. 1 : 50 000 scale), this is not often the case for small-scale mapping which is drawn from multiple sources and often not for thematic mapping at any scale. Certainly there are no published standards of accuracy for many of the maps used in the CORINE Programme, and it is doubtful whether they would meet the US standards for map accuracy (see Rhind and Clark 1988). Thus the user may find it difficult to judge the relationship between a line on the map and a line on the ground; in practice, it is doubtful whether the positional accuracy of much of the material within the development database is greater than 1 km. The representation of 'fuzzy features' such as soil boundaries is inherently less accurate than is that of physically discrete ones like railways (see Burrough 1991 in this volume).

Distortions in the position of features stored in the database can also be introduced inadvertently through data processing. This may be extremely obvious (for instance, digitizing spikes), or alternatively very subtle and not immediately apparent. A good example of the latter is provided by the soils data within the interim CORINE database, derived from the already published paper EC soils map (Tavernier 1985). This original map was compiled from a set of national soil map sheets which, although each was on a known projection, were only minimally transformed in conversion to the whole EC map. Perhaps the most severe effects of this method of compilation were the distortions in soil boundaries along the edges of the original, national sheets in order that a continent-wide continuous paper map could be assembled for wall display purposes. Worse still, the map was then partitioned into other map sheets for publication purposes. Described thus, the compilation process seems to have been inept but it must be remembered that no thought whatever had been given to the final map being anything other than a free-standing pictorial display when its production began in the late 1970s.

The digital representation of the EC map was produced through scanning and subsequent vectorization of the final, published (and internally distorted) map sheets, thus embodying the original map's significant distortions along original (but, by then, unrecognizable and unrecorded) sheet edges. These distortions were only discovered when the soils and other data sets – supposedly derived from the same ONC topographic map base (Rhind and Clark 1988) – were overlayed. To allow integration with other data sets, the distortions had to be removed. The first attempt to solve the problem involved the use of 'rubber sheeting' with over 7000 control points. Unhappily, this proved unsuccessful because of the lack of control points available on both the spatial template (the ONC map) and the soils database in certain critical areas, allied to the nature of the errors. To achieve the desired result, substantial 'detective work' had to be carried out to discover the processes through which the maps had been. After that, the digital data were divided into the original sheets and reprojected to their known projections. The distortion introduced through the original edge-matching had then to be removed by 'rubber-sheeting' techniques; the data set now overlays the topographic coverage, but cannot be compared directly with the source document from which the digitizing was carried out.

Attribute accuracy is a separate issue, and defines the closeness of the attribute values to their true values. Gross errors (such as miscoded polygons) may become obvious through use of the

data and familiarity with the area or comparison against other sources; if so, these are easily fixed. Other errors are more subtle. Problems in categorizing what are in reality continuous variables (e.g. soil properties) may be compounded by the difficulty in defining the position of the polygon on the ground. Finally, although the interim CORINE database aims to avoid the storage of derived data, in some cases this is unavoidable. This can give rise to the derivation of indices by differing methodologies (e.g. five different formulae were found to be in use between the 12 national climatological organizations for the derivation of the monthly maximum daily temperature, and eight for the calculation of potential evapo-transpiration statistics).

The second component of data quality is that of completeness. This can also be expressed in terms of position and attributes. In the former, parts of the data set may be missing; for instance, when the CORINE project began, a digital cartographic database of topography at a scale of 1 : 500 000 was made available by the Ministry of Defence in the United Kingdom. This data set avoids many of the difficulties with the IfAG data set referred to above and also contains contour information. Unhappily, it is complete only as far south as 46° N, thus excluding much of the Mediterranean area – one of the key areas for study in the Programme. The mismatch between data extent and the area of EC needs reflected the differing responsibilities of the two organizations involved.

In the case of point-sampled data sets, the concept of positional completeness is more difficult to determine; there is a need to ensure consistent and representative density of sites across the study area. There are no invariable rules for determining this; Burrough (1986) shows that the sampling density for boulder clay (which varies widely across short distances) should be much higher than that for sandstone (which is generally far more consistent in its properties). In climatic data sets, more sites are needed to represent rainfall accurately (which is locally distributed) than solar radiation (which is more regional in character). In essence, therefore, sampling should be related to autocorrelation in the data set. In practice, this is rarely known before data are collected and sampling strategies are often complicated by pragmatic and even political considerations.

Such problems in data collection also determine the completeness of attributes, both through space and time. Many data sets within the interim CORINE database are temporally incomplete, due either to failure of recording equipment or to disruption to the monitoring systems (for instance during the period 1939 to 1945). Positional information is recorded for some data sets (for instance, that on biotopes), but sometimes lacks a full range of attribute information (e.g. species at that site). As indicated earlier, problems of data quality become particularly acute (indeed, they may only be recognized) when data sets are overlaid during environmental modelling. Users should be aware of the limitations that map source scale places on this process; data derived from small-scale sources cannot realistically be used in conjunction with that collected at larger scales because of the effect of scale on accuracy and spatial precision. For example, the CORINE database contains data on land cover compiled at a scale of 1 : 100 000 and derived from the interpretation of Landsat MSS satellite imagery. For studies on land use in the Mediterranean region, this needs to be overlaid on to the soils map derived from manual compilations of pre-existing national soils maps at 1 : 1 million scale. Clearly, the former should be generalized before this operation can take place. While a tenfold linear generalization will lead to a (possibly unacceptable) loss of information in the land cover data set, the alternative of expanding the soils data set to 1 : 100 000 is unacceptable as it would simply magnify the distortion already inherent in that data set.

A related problem is that of the spatial relationships between different data sets. Often these are hidden or implied but, unless they are known, then the user is at risk of drawing conclusions from the analysis which are at best tautologous or, at worst, nonsense. The problem is particularly acute where data sets are thought to share common boundaries (e.g. soil and vegetation which may terminate along river banks). If the two data sets are derived from different sources (e.g. maps of differing projections and scales) then, when overlaid, the boundaries may no longer be coincident and sliver polygons will occur. The problem for the user is to decide whether these are real or whether they are simply a reflection of variations in data quality. To answer this, the user requires specialist knowledge of the data sets,

including the history of their derivation; without this, they may collapse polygons which in reality are discrete and real and thus force a spurious correlation between the data sets.

Database volume and update

Because the CORINE database is still under development, a range of problems have been identified which have still to be addressed. These include the question of handling large data volumes as the spatial resolution increases and the maintenance and update of the database. The CORINE database, at the time of writing, totals about 750 megabytes when held in ARC/INFO format. While small by global standards (especially in comparison with those derived from remote sensing imagery; see Clark *et al.*, 1991; Townshend 1991 in this volume), it is expected to increase considerably in the future. Many environmental processes operate at resolutions considerably finer than 1 km^2 (the best attainable resolution on the ground of the existing database at a scale of 1 : 1 million and very unlikely to be attained consistently). Thus an increase in resolution is clearly desirable and, indeed, essential if the database is to be put to routine practical use for many purposes. Indeed, such an increase in resolution is already reflected in data holdings on both land cover and coastal erosion which were compiled at larger scales. The NATO requirements for digital topographical cover across Europe at 1 : 50 000 scale ensure that, even in the medium term, the potential size of the EC database is measured in terms of gigabytes rather than megabytes.

Databases of this size require careful design and structuring if the information within them is to be readily accessible to the user; it is already clear that some form of spatial partitioning ('tiling') of the CORINE database is required, whether achieved by system designer or internally by the system itself. Originally stored as one seamless whole, this has the advantage of simplicity of database design but increases processing time for user access to only part of the area of the European Community. Some early experiments to identify an optimum tiling strategy using ARC/INFO suggested that partitioning by country would be most appropriate and readily understandable by the end-user (Wiggins 1986); many queries arose on a country-by-country basis. However, data volumes by tile were still too large to give acceptable access times. An alternative series of tiles of 2 degrees longitude by 1 degree latitude (which happened to be the 'building blocks' of one of the sets of map sheets used) were constructed; while improving the access times, the pattern was not readily identifiable by the user. It is now clear that patterns of user access to data should determine the tiling structure; irrevocable decisions on partitioning have been deferred until there is more extensive use of the database by a variety of different users and until selection is made of the final system to be used.

The problem of updating the CORINE database has yet to be put to the practical test; clearly there should be procedures for any database to ensure that its contents are accurate and up to date. The currency of data and the frequency of their update are a function of the type of data and of the uses of them. In the case of environmental databases, many update cycles are quite long; geology and soils, for instance, change most rapidly through re-interpretation rather than through natural processes! Hence replacement of a whole data set or aggregation or disaggregation of classes in the data are normally required with such data. Meteorological data, at least when stored as 30 year means, are also fairly stable. In contrast, both biological populations and patterns of land use and cover may change extremely rapidly; their update cycles are thus much shorter. Procedures for updating have yet to be determined for the CORINE database; what is already clear is that revision will be as resource-intensive as was compilation of the original data. A further complication is that the responsibility for revision of primary data will rest with the data suppliers (many of whom are national agencies in the member states of the EC) rather than the users or the data holders (the EC). In practice, therefore, updating of such a multinational database as CORINE is likely to require much collaboration at the political as well as technical levels and may require EC directives.

Database documentation

Standard procedures for database documentation must be established if the user is to know the history of each data set and thus have some understanding of its quality and potential for use. There are several levels of documentation: of individual features or of classes of features (or variables) within data sets, and of the data sets themselves. Chrisman (1984,

1991) has argued for the inclusion of information on data quality and reliability within each data set by individual feature. Further, the system should be intelligent enough to act upon this information, in order to guard against misuse. While obviously this is one ultimate aim of the CORINE database, the information on which to judge data quality at present is often unavailable, is often only in free-text form where it does exist and its inclusion would have some implications for data volumes.

Documentation of standards for data collection and definition will go some way to ensuring an increase in data quality and attribute accuracy. The CORINE Programme team is presently compiling a catalogue of data definitions. It would be advantageous to construct and disseminate these before data collection takes place, in order to ensure more rigorous selection and thus increase data quality. In practice, many are at present established either during or after data collection, but are still useful in identifying errors in the database (for instance the inconsistencies of definition within the climate data sets noted earlier). Standard procedures for the documentation of the history of each data set have also been established for the CORINE database. In this way, users are able to determine the source of each data set and follow its history of processing and assimilation into the database. The history files and audit trail facilities available in some GIS are invaluable in this respect.

ISSUES OF USER ACCESS TO ENVIRONMENTAL DATABASES

Underlying the various issues concerning environmental data and GIS technology to handle it are questions concerning the organizational background: where should the database be sited, to whom should it be accessible and for what purposes?

Centralized versus distributed database

The CORINE database is presently centralized at one site, but this need not be a model for other environmental databases or even for CORINE in the longer term. There are three possible scenarios: all data at a central site; all data at many sites; or some data across a range of sites with a greater or lesser degree of transparency in access to the user.

Significant improvements in networking and communications technology over the past decade have provided direct access for users to many centralized databases. The idea of many users having access to environmental data distributed across many databases is not yet, however, as realistic for a number of reasons. The requirement for multiple variables across large geographical areas can result in massive volumes of data for file transfer, and this may be complicated by the difficulties in processing some typical GIS operations over a network (particularly when complex graphics are involved) and the inexperience of many users in use of network technology. An alternative is to distribute the database on optical storage media for local access but this in turn raises problems of database update. The ever-decreasing cost of storage of data on CD-ROM and the economic possibility of repeat pressing at intervals may resolve this issue. Another alternative is that developments in data broadcast offer realistic longer-term prospects.

User access

Free and uncontrolled access to a centralized CORINE database is a technical possibility. In reality, however, it is at present neither feasible nor desirable. Many users have only limited knowledge of the operation of a GIS and, without significant improvements in the user interface and/or user education, this is likely to pose a practical barrier to free access to the data. Other users, while technically capable of accessing the database, have only limited understanding of some of the issues of data quality noted above. Though it may be argued that it is not the duty of the database builder to prevent *bona fide* users from misusing the data, there are strong scientific, ethical and political reasons for doing so. For instance, many environmental issues are scientifically and politically very sensitive; misinterpretation of the data and results of analyses could lead either to the establishment of inappropriate policies or to the discrediting of the whole information system, providing a justification for suspending its implementation.

The CORINE database is accessed at present through a user service; access to the data is via in-house, 'expert' users only. This allows use of the data to be carefully regulated and inappropriate uses filtered out. It also offers the opportunity for education through discussion with users of the design of any data analysis or output, and the provision of advice on the most appropriate analytical techniques. A disadvantage is that it may deter use of the database or slow down access. More seriously, if not sensitively and openly implemented, it may amount to a form of data censorship, filtering out politically or administratively undesirable queries.

In the long term, more open use of environmental databases may be achieved through the use of expert systems, with their own built-in rules for data use. Unfortunately, while examples of such systems have been demonstrated (see, e.g. Smith and Ye Jiang (1991 in this volume) and Smith, MacKenzie and Stanton (1988) on the development of an expert system to support zoning of the Australian Great Barrier Reef), they are still some way from widespread operation and the rules which they can apply are only as good as the people who devise them. In the case of the CORINE system, this presents serious difficulties for the database is not yet in a sufficiently stable state nor is the management science yet sufficiently advanced to permit the application of the 'hard and fast' rules required of most expert systems. In particular, the user needs are not yet sufficiently understood to define the rules and the complex interaction between environmental variables remains inadequately understood.

FUTURE DEVELOPMENT

Although still in its formative stages, the CORINE database already represents one of the most substantial fully integrated systems in the world – certainly if those comprised wholly of remote sensing imagery are excluded. Already, there have been three positive achievements (Wyatt, Briggs and Mounsey 1988):

- some harmonization in existing practices and the acceptance of standards for recording environmental data;
- a demonstration of the feasibility of establishing one centralized database to meet the requirements of a diverse variety of end-users; and
- the development of similar integrative activities at national level which, in themselves, reinforce the improvement in data collection practices.

But there are also some lessons which should be taken forward into the next stage. First, it should be noted that the issues behind the development of environmental databases are largely non-technological; indeed the rate of development of technology is (at least at present) outstripping both data quality improvements and the ability of the user to operate it. Substantial efforts in user education are needed to resolve this issue.

Secondly, the development of an environmental database needs to be well resourced. Four years of data collection and integration merely confirms that such database creation is an expensive process, principally because tasks which appear conceptually simple are either highly labour intensive or more complex (and thus time consuming) than originally envisaged. This is especially the case when judged by those without practical GIS experience.

Thirdly, the CORINE database was developed as a reaction to existing problems of nature conservation, acid deposition and conflicts of land use in the Mediterranean. But, to be most effective, the creation of environmental databases should be pro-active, backed up with sufficient resources to involve modellers as well as database builders. The gestation period for assembling environmental databases is such that only by early – and, ideally, prior – identification of the key processes which govern environmental change can databases be developed which make real contributions at the most apposite moment in the battle against environmental degradation.

Fourthly, environmental databases should be built through better efforts on overall system design and basic data requirements, rather than through the random provision of information by disparate policy themes. Although three specific topics for study were defined at the outset of the CORINE Programme, little initial thought was given to definition of the fundamental data requirements which should ideally underpin an environmental

database. The *ad hoc* approach adopted by the CORINE Programme is probably not untypical of 'first-generation' systems driven by enthusiasm and a need to demonstrate results, rather than having a commitment to longevity and sound principles of design. This should not be a long-term policy.

Finally, the development of any environmental database requires full and substantial organizational support by all interested parties. In the case of CORINE, these include directorates within the European Community itself, national governments and their agencies and international organizations. The international dimension of environmental problems requires reliable information and rigorous analysis. But, as Wyatt *et al.* (1988) have noted, the balancing of political objectives and financial commitment against technical reality and scientific rigour is one of the most elusive goals in policy making. It would be unrealistic to assume that science alone will ever dictate how, when and why environmental GIS and their databases are created and used.

REFERENCES

Benyon D (1990) *Information and Data Modelling*. Blackwell Scientific Publications, Oxford

Blatchford R P, Rhind D W (1989) The ideal mapping system. In: Rhind D W, Taylor D R F (eds.) *Cartography Past, Present and Future*. Elsevier, London, pp. 157–68

Briggs D J, Mounsey H M (1989) Integrating land resource data into a European geographical information system. *Applied Geography* **9** (1): 5–20

Burrough P A (1986) *Principles of Geographical Information Systems for Land Resource Assessment*. Clarendon Press, Oxford

Burrough P A (1991) Soil information systems. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 153–69, Vol 2

Calkins H W (1991) GIS and public policy. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 233–45, Vol 2

Canada Department of Forestry and Rural Development (1965) *The Canada Land Inventory: objectives, scope and organisation*. Report No. 1. Ottawa, Canada Land Inventory

Chrisman N R (1984) The role of information quality in a GIS. *Cartographica* **21** (2/3): 79–87

Chrisman N R (1991) The error component in spatial data. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 165–74, Vol 1

Clark D M, Hastings D A, Kineman J J (1991) Global databases and their implications for GIS. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 217–31, Vol 2

Clarke A L (1991) GIS specification, evaluation and implementation. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 477–88, Vol 1

CEC (1990) CORINE: *Examples of the Use of the Results of the Programme 1985–90*. Directorate General of the Environment, Nuclear Safety and Civil Protection, Commission of the European Communities, Brussels

Department of the Environment (DoE) (1987) *Handling Geographic Information. Report of the Committee of Inquiry chaired by Lord Chorley*. HMSO, London

Didier M (1990) *Utilité et valeur de l'Information Géographique*. CNIG Economica, Paris

EPA (1987) *Sharing Data for Environmental Results*. State/EPA Data Management Program Project Report 1987. United States Environmental Protection Agency, Washington

Epstein E F (1991) Legal aspects of GIS. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 489–502, Vol 1

Fisher P F (1991) Spatial data sources and data problems. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 175–89, Vol 1

Flowerdew R (1991) Spatial data integration. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 375–87, Vol 1

Guptill S C (1991) Spatial data exchange and standardization. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 515–30, Vol 1

Maguire D J, Dangermond J (1991) The functionality of GIS. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 319–35, Vol 1

Mott J (1990) The National Resource Information Centre – data directory, data broker. In: Parvey C, Grainger K (eds.) *A national Geographic Information System – an achievable objective?* AURISA Monograph 4. AURISA, Eastwood New South Wales, pp. 57–60

Muller J-C (1991) Generalization of spatial databases. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 457–75, Vol 1

Official Journal of the European Community (1985) Council Decision on 27 June 1985 on the adoption of the Commission work programme concerning an experimental project for gathering, coordinating and ensuring the consistency of information on the state of the environment and natural resources in the Community. OJ L 176, 6 July 1985

Rhind D W, Clarke P K (1988) Cartographic inputs to global databases. In: Mounsey H M (ed.) *Building Databases for Global Science*. Taylor & Francis, London, pp. 79–104

Rhind D W, Wyatt B K, Briggs D J, Wiggins J C (1986) The creation of an environmental information system for the European Community. *Nachrichten aus dem Karten und Vermessungswesen Series 2*, **44**: 147–57

Smith J L, Mackenzie H G, Stanton R B (1988) A knowledge-based decision support for environmental planning. *Proceedings of the 3rd International Symposium on Spatial Data Handling*. International Geographical Union, Columbus Ohio, pp. 307–20

Smith T R, Ye Jiang (1991) Knowledge-based approaches in GIS. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 413–25, Vol 1

Tavernier R (1985) *Soil Map of the European Communities. 1 : 1 000 000*. Office for Official Publications of the European Communities, Luxembourg

Tomlinson R F, Calkins H W, Marble D F (1976) *Computer Handling of Geographical Data*. UNESCO, Paris

Townshend J R G (1991) Environmental databases and GIS. In: Maguire D J, Goodchild M F, Rhind D W (eds.) *Geographical Information Systems: principles and applications*. Longman, London, pp. 201–16, Vol 2

Tsichritzis D C, Lochovsky F H (1982) *Data Models*. Prentice-Hall, New York

Whimbrel Consultants Ltd (1989) *CORINE Database Manual, Version 2.1*. Brussels

Wiggins J C (1986) Performance considerations in the design of a map library: a user perspective. *Proceedings of the ARC/INFO Users' Conference*. ESRI, Redlands California

Wiggins J C, Hartley R P, Higgins M J, Whittaker R J (1987) Computing aspects of a large geographic information system for the European Community. *International Journal of Geographical Information Systems* **1** (1): 77–87

Wyatt B K, Briggs D J, Mounsey H M (1988) CORINE: An information system on the state of the environment in the European Community. In: Mounsey H M, Tomlinson R F (eds.) *Building Databases for Global Science*. Taylor & Francis, London, pp. 378–96