A data management strategy for scientific research.

A. M. Roberts
Institute of Hydrology
Wallingford, OX10 8BB, UK
Tel:+44(0)1491692363 Fax:+44(0)1491692424 E-mail: amr@ua.nwl.ac.uk

Abstract

Since the mid 1980's, political and technological changes in UK scientific research have heightened awareness of the economic value and importance to scientific research of quality datasets and Environmental Information Systems (EIS). As a result of these changes, the UK Natural Environment Research Council (NERC) has funded more interdisciplinary Thematic Programmes. These programmes have highlighted the need for a coherent data management policy so that maximum long term gain is achieved by the science community. One such programme is the Land Ocean Interaction Study (LOIS). This is a £30 million Programme that involves researchers from both UK public sector Institutes and Universities.

This paper discusses the LOIS data management strategy, its problems and its solutions in the large scale data collection and storage that underpins the multi-disciplinary modelling efforts. The establishment of a Data Centre concept has enabled scientists to access, via computer networks, integrated datasets from the air, terrestrial, freshwater, geological and marine phases of the environment.

Keywords

Data integration; data management; science research.

INTRODUCTION

At a workshop on Rural Information for Forward Planning in 1986, it was suggested participants consider the view that, the modern state needed a basic data infrastructure (Haines-Young, 1988); it is my opinion that this remains true more than ten years later. Developers of EIS have several crucial issues to address. First, there is a pressing need to create databases and links to appropriate analytical models to create Decision Support Systems. These systems

Environmental Software Systems Vol.2 R. Denzer, D.A. Swayne & G. Schimak (Eds.)

will meet the needs of regulators and policy makers in the future. Second, a higher priority should be given to the issues of data awareness, access, integration and quality. (Strachan et al, 1996).

Since the mid 1980's there have been fundamental changes in the perception of how important good quality databases are to scientific research. These changes have been reflected in Government policies, organisational structures in UK science and in EIS technology. In a UK government policy document, it was concluded that, the monitoring of environmental trends is an area of science in which government must necessarily take the lead, and requires long-term commitment and financial support. The implication of this policy has ensured that the implementation of a long term data management policy in UK science now has a high priority, albeit with delay in concomitant funding.

In response to these recent scientific and political changes, NERC has carried out extensive changes in the organisation of its research. To address the generic problems of data acquisition, data management, data use and charging for data, NERC has implemented a Data Policy Plan (NERC, 1996). This Plan has raised awareness of the imporatnce of data and data systems in British science. Due to Government policy pressures, it has become increasingly important for NERC to work across traditional discipline boundaries in order to align its strategic initiatives with the end user. NERC's Thematic Programmes are achieving this, by bringing together scientists from public-funded Institutes and Universities to address cross disciplinary research areas.

The aim of this paper is to describe how the issues of data awareness, access, integration and quality have been addressed in a major NERC Thematic Programme. In addition, the EIS used in one of the Programme's Data Centres, is used as a case study to illustrate how atmospheric, marine, terrestrial and freshwater data have been integrated and stored for use by the Programme's scientists.

THE LAND OCEAN INTERACTION STUDY (LOIS)

LOIS is NERC's current £30 million flagship Thematic Programme addressing the effect of increasing exploitation of the coastal zones in the UK. This exploitation, if not carefully managed, could result in the degradation of a sensitive environment. The objective of LOIS is to extend the knowledge of the environmental fluxes between the air, land, ocean continental shelf and the ocean shelf edge.

LOIS has two main study areas. The main area covers the river basins from Berwick-on-Tweed in south east Scotland to Great Yarmouth in south east England and has a great diversity of land uses and river types. The northern uplands have peat moorlands with relatively unpolluted streams. Large industrialised conurbations, with rivers subject to high effluent inputs, occupy the central part of the study area. Low lying fens, with intensive agriculture and slow-moving nutrient-rich rivers, are the characteristics of the remaining south eastern part of the region. As the North Sea, which borders this study area, does not have a continental shelf boundary, the study area for the LOIS shelf edge component is situated on the western side of the UK, at the edge of the continental shelf south west of the Hebrides.

DATA MANAGEMENT STRATEGY

Kennedy and Guinn (1975) describe information systems as essentially data driven: 'While models which use the data are important to support the decision-making activities of those who use the system, a large portion of the investment will be obtaining, converting and storing new data.' This was still true twenty years later when LOIS was being established, and therefore to contain costs and effort, an organisational structure servicing the needs of the research community was initiated.

The data management problem

With the problem of providing scientists easy access to the datasets they required in mind, a new approach to the management of data was established for Thematic Programmes. Lowry et al in their papers examine the problems of data management and quality control to support Thematic Programmes in the marine sciences (Lowry & Cramer, 1995; Lowry & Loch, 1995). They found that the scale and cost of the science involved had raised the awareness that the datasets produced are a valuable long term resource and should not be allowed to remain solely in the hands of individual scientists. To justify their high costs, Thematic Programmes must produce 'deliverables'. In science these usually take the form of published papers, but a further deliverable is now a clearly identified Programme dataset.

The central aim of LOIS is to study the environment of the land-ocean boundary as an integrated whole; therefore, models and analyses require simultaneous access to many different data types. Before the start of LOIS, these

data access would have been difficult as they were held in a variety of formats, often on incompatible hardware systems.

The data management solution

To address these problems and to reduce the time spent on data acquisition by scientists, the LOIS management team established a Data Centre for each of the Programme's components. Each Data Centre had the same mission to: identify requirements; set standards; acquire/collate/store data and finally, to disseminate these data. This organisational structure has achieved its aims of providing easy access to cross disciplinary data, despite a wide geographical spread of the Centres. This success has been accomplished by the high degree of coordination within the LOIS DATA management team. This team has one or more representatives from each of the Data Centres as well as a representative from each of the Programme's scientific components. This mix of representation has encouraged the scientists to seriously think about data planning issues and data managers to consider the scientific issues. Thereby a science/data symbiosis has been maintained.

THE INSTITUTE OF HYDROLOGY (IH) LOIS DATA CENTRE

The IH LOIS Data Centre was established to support the needs of the river basin modellers within the Rivers component of LOIS. The river quality and river quantity data from the whole of the LOIS study area were required by the scientists. These data were held by individual regions of the Environment Agency (EA) in England and Scotland and had no standard formats or coding systems. These data have now been unified and standardised for all regions of the LOIS study area. As a result, this database is now the single largest integrated river database in the UK.

IH LOIS Data Centre Environmental Information System data model

The EIS used by the IH LOIS Data Centre is a version of the Water Information System (WIS). This system was originally designed and developed at the Institute of Hydrology to store and manage all river basin data. In the literature, many descriptions for spatial data models can be found (eg Sussman, 1993; Hadzilacos and Tryfona, 1996; Tang et al, 1996), but WIS is atypical, in that it does not distinguish between spatial and time series attributes. WIS allows all attributes to be stored in a single unified database (Hill, 1996). This enables the user to record the history of any object or feature, as it moves through space and time.

For example, a river water quality monitoring site would be classified as a 'feature' and the 'attributes' which describe, or are observed at, this site may include many variables, eg. a site name, location, unique reference number, river flow, pH, suspended sediment etc.

Although WIS was developed for river basin data, due to its flexible data model design, it has been shown that data from any of the LOIS components (eg marine or atmospheric data) can be stored and retrieved. The logical idea of this design is a cube where the axes represent features, attributes and time. Physically, the data are held in tables in ORACLE (see Figure 1) which is a commercially available Relational Database Management System.

FID	DID	Date/time	Value	Qualifier	Method	Status

FID	The identifier of the feature to which the value belongs
DID	The dictionary identifier of the attribute to which the value belongs
Date/time	The date and time at which the value was measured
Value	The data value. This can be either a spatial or time series value
Qualifier	A qualifying code eg <, >, = etc.
Method	A code of the attribute determination method
Status	A code to indicate any validation procedures applied to the value

Figure 1 The generic physical table design for WIS

Data integration

To achieve the integration of time series and spatial data within a single database, the IH LOIS Data Centre was faced with the challenge of matching data types from different sources. The problem being that no standard coding system exists, for similar data types from the various regions of the EA, or between water quality data from freshwater and marine sources.

Smith et al (1987) in their paper, and from our own experience in the Data Centre, have found that data for input to a database are typically acquired in a variety of forms. These forms may include graphical data, non-spatial information (i.e. numerical data, descriptive or attribute data and textual data) from both printed and digital files. WIS requires data to be in a specific format for importing data. Therefore, there was an initial task to write a suite of

programs that would convert the data from the 18 different formats in which the data was supplied to the Data Centre.

Access to the database

Having solved the problem of how to integrate and store LOIS data in an EIS using WIS, the mechanisms for making the data available to scientists then had to be considered. The IH LOIS Data Centre set up five methods of access:

- The WIS geo-referenced database interface can be used to browse the data as maps, graphs or reports.
- Query the tables direct using the Structured Query Language (SQL). Users may also write their own programs in PRO*C or PRO*FORTRAN which has embedded SQL.
- Open Data Base Connectivity enables PC applications to access the database. This method of database access allows the presentation of data in applications such as Microsoft Excel and Access.
- Data can be exported from the database on various forms of computer media.
- A new form of database access, using the latest form of Client/Server technology is currently being developed and tested at the IH LOIS Data Centre. This technology will enable a user to access the database remotely, via the World Wide Web (WWW) using Internet services, and download data to the users' web page.

CONCLUDING REMARKS

In this paper, the strategy and problems of how to provide integrated scientific data from a EIS has been discussed. The strategy has been successful despite the fact that nothing on this scale has been undertaken before in UK environmental science. The benefits to the research scientists of the LOIS Data Centre policy are:

- Data from different scientific disciplines have been integrated successfully.
- New field and model data are being conserved for future generations of scientists in an active form.
- Advances in database query technology are improving efficiency of access to environmental results.

There have been problems encountered while setting up the IH LOIS Data Centre. The chief of these has been in acquiring or distributing data sets, from or

to, third parties. The delay in acquisition has highlighted the need for simpler Copyright and Intellectual Property Rights arrangements with the agencies, especially those in or close to the public sector, supplying data.

REFERENCES

Hadzilacos, T. and Tryfona, N. (1996) Logical data modelling for geographical applications. *Journal of Geographical Information Systems*, **10**, 2, 179-203.

Haines-Young, R.H. (1988) Data needs for rural planning. In *Rural Information for Forward Planning*, (ed. R.G.H. Bunce and C.J.Barr) Institute of Terrestrial Ecology Symposium No 21, Cumbria, 110-112.

Hill, D.R. (1996) Search Mechanisms for querying the time dimension in 4D-GIS. *Proceedings of the 1st International Conference on Geo-Computation*, University of Leeds, UK, Sept 1996.

Kennedy, M., and Guinn, C. (1975) Automated spatial data information systems: Avoiding failure. Urban Studies Center, Louisville, p76.

Lowry, R.K. and Cramer, R.N. (1995) Database applications supporting Data Management, *Geological Society Special Publication*, **97**, 103-107.

Lowry, R.K. and Loch, S.G. (1995) Transfer and SERPLO: powerful data quality control tools developed by the British Oceanographic Data Centre. *Geological Society Special Publication*, **97**, 109-115.

Natural Environmental Research Council (1996) NERC Data Policy Handbook, January 1996, NERC, Swindon.

Smith, T.R., Menon, S., Star, J.L. and Estes, J.E. (1987) Requirements and principals for the implementation and construction of large-scale geographic information systems. *Journal of Geographical Information Systems*, **1**, 1,13-31.

Sussman, R. (1993) Municipal GIS and the enterprise model. *Journal of Geographical Information Systems*, 7, 4, 367-77.

Strachan, A.J. and Stuart, N. (1996) UK Developments in Environmental GIS. *Journal of Geographical Information Systems*, **10**, 1, 17-20.

Tang, A.Y., Adams, T.M. and Usery, E.L. (1996) A spatial data model design or feature-based geographical information systems. *Journal of Geographical Information Systems*, **10**, 5, 643-659.

Anne M Roberts. I have been a computer scientist at the UK Institute of Hydrology (IH) for 19 years. In the past I have been responsible for designing and implementing hydrological databases. At present, I am the Head of the IH Software Operations Section, responsible for the ongoing development and maintenance of IH software packages.