Org. Divers. Evol. 3, 207–213 (2003)

http://www.urbanfischer.de/journals/ode

ORGANISMS DIVERSITY & EVOLUTION

SOMBASE – Southern Ocean Mollusc Database: A tool for biogeographic analysis in diversity and ecology

Huw J. Griffiths*, Katrin Linse, J. Alistair Crame

British Antarctic Survey, NERC, Cambridge, UK

Received 30 September 2002 · Accepted 4 April 2003

Abstract

Databases and Geographical Information Systems are becoming increasingly popular tools for biogeographic analysis. The reliability of the analysis depends on the accuracy of the entered data and the ability to add changes in systematics and taxonomy. The main aim of the Southern Ocean Mollusc Database (SOMBASE) project is to set up a database linked with a GIS and user-friendly front end. This system allows for the inclusion of other taxa in the future. Currently the database consists of records for 950 shelled gastropod and 136 bivalve species and 2800 sites from the Southern Ocean.

The database contains fields including: 1) Taxonomic authorship, synonyms, higher level classifications, diagnostic morphological characters, and ecological features; 2) Distributional information including substrate and depth; 3) Bibliographic information.

The maps can display selected sites based on any query of any field or combination of fields in the database. An online version of the database is available with distribution maps for all taxonomic levels (www.antarctica.ac.uk/SOMBASE).

Key words: GIS, database, biogeography, distribution, systematics, taxonomy

See also Electronic Supplement at **http://www.senckenberg.de/odes/03-12.htm**

Introduction

In recent years relational databases have started to displace faunal lists in biogeographical and diversity studies (e.g. Budd et al. 2001, Rosenberg 1993, Hill et al. 2000, Zhang et al. 2000). Whilst faunal lists can demonstrate the distribution of large numbers of taxa, large-scale and global perspectives on biodiversity are often lacking, as well as information on community structure, abundance, ecology and environment (Grassle & Stocks 1999). These data are of importance to analyse and explain diversity patterns and to test ecological and zoogeographic hypotheses. Furthermore, it is becoming apparent that when relational databases are linked to a Geographical Information System (GIS) they become an even more powerful tool for taking on large-scale biogeographical patterns (Markwick 2002, Markwick & Lupia 2002).

We aimed to create a comprehensive database of Southern Ocean marine invertebrate taxa to test the following hypotheses:

1) Are there latitudinal and longitudinal gradients in taxonomic diversity?

2) Is local species richness a reflection of regional species richness?

3) Do regional trends exist in functional groups?

The database has been developed using shelled gastropods and bivalves, two of the most common and taxonomically well-studied groups in the Antarctic region (e.g. Gutt et al. 2000, and references therein).

In this paper we demonstrate how this database has been constructed, and some of its practical applications. An online demonstration of distribution maps for all taxonomic levels within SOMBASE is available (www.antarctica.ac.uk/SOMBASE).

Material and methods

Design and set-up of a relational database

Relational databases are a way of organising and storing data in tabular form through a system of relation-

***Corresponding author:** Huw J. Griffiths, British Antarctic Survey, NERC, High Cross, Madingley Road, Cambridge CB3 0ET, UK; e-mail: hjg@bas.ac.uk

ships based on shared information. The use of a relational database enables the division of data into logical categories and hierarchies, thus reducing data repetition and increasing query speed and capabilities. SOMBASE is a biogeographic (biological information within a spatial context) database, which relates species information to data on the localities at which the species were found. The design of SOMBASE enables users to query and search its contents based on species attributes, site attributes, or a combination of attributes from species and sites.

The initial stage of database design involves identifying the user requirements of the database, i.e. what information is required by the user and at what level of detail, and how this information will be used. Once the aims of the database are established it is important to assess the availability, usefulness, search ability, and format of the data to be entered. It is advisable to review the available data to create a list of standardised terms or descriptions. This is to ensure that the required data is selected in a query and not ignored because it has a different description, although it means the same thing.

In the case of SOMBASE the data are divided into taxonomic and locality information (Fig. 1). The taxonomic data are further categorised by taxonomic level, i.e. "Family", "Genus" and "Species". In a similar way the locality data are split into "Zone" (large areas of the globe such as Antarctica, South America, the Southern Pacific, etc.), "Area" (regions within a zone such as island groups, seas or coastlines), and "Site" (a point or tow sample location with specific latitudes and longitudes). A table is created for each level of the data. The tables are then related by shared information, common fields. These relationships are dependant on the fact that, in at least one of the tables, the field involved in the relationship contains unique values, i.e. a key field. In the case of a one-to-many relationship, the key field in the higher-level table, e.g. the "family" field in the "Family" table, is related to many records in the next-level table by the matching field (foreign key), e.g. the "family" field in the "Genus" table (Fig. 1).

Due to the fact that site and species names often occur more than once, the tables do not depend on these names as a unique identity (key fields). Instead we adopted ID codes, which are created using an autonumbering feature with the site IDs being a combination of alphanumeric criteria taken from the data source used.

Problems arise when relating the species data to the locality data. It is possible that a species may be found at more than one locality and a site may have many species present; this is known as a many-to-many relationship. To resolve the complications arising from a many-to-many relationship an intermediate joining table can be used. In the case of SOMBASE the "Find" table serves this purpose (Fig. 1). It contains a find record for every occurrence of an individual species at a site. This breaks down the many-to-many-relationship into multiple one-to-many relationships. This allows each species record to be linked to one or more sites, and also the sites to be associated with one or more species. In this way, several hundred site records and several hundred species records can amount to several thousand find records.

Repetitive information such as family or genus names is only entered once, and additions in other tables are re-

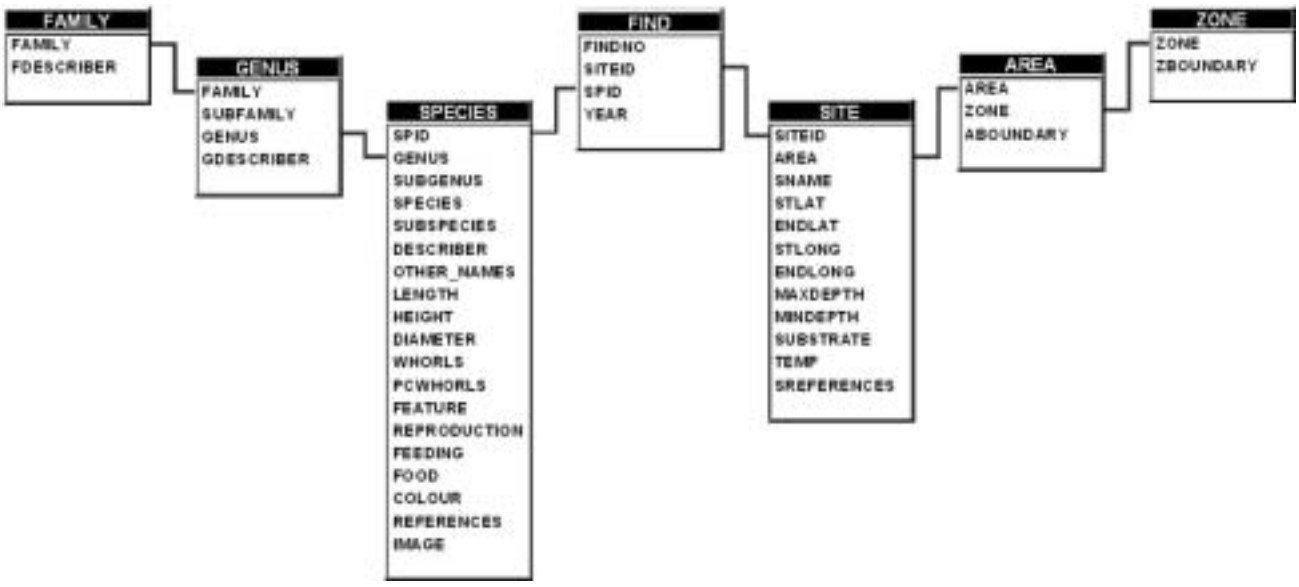

Fig. 1. Design of the database relationships, showing the different tables, key fields and "one-to-many" relationships.

lated to this original entry. The higher-level tables (e.g. “Family”, “Genus”, “Zone”, “Area”) hold mainly definitions and taxonomic references. Further information on specific taxon or locality records can be added into the species and site tables. The other fields in the “Species” table have been selected for their systematic, taxonomic and ecological value in the search for biogeographic patterns (Fig. 1). The fields in the “Site” table are selected to describe the geographical and environmental factors that affect the species occurrence.

Fields are defined by simple criteria; e.g. text or numeric, length of text string, or number of decimal places. Key fields and essential data, such as latitudes and longitudes in the site table, can be defined as ‘required’, i.e. null values are not allowed. The use of indexes on key search fields increases the speed of queries by organising the data automatically in advance. Indexes can be particularly useful when working with large datasets if used correctly on small numbers of fields. The over-use of indexes can actually slow down the performance of a database.

Software and hardware

SOMBASE consists of a main datastore, a database front end, and a GIS (Fig. 2). When creating a set-up similar to SOMBASE the choice of database software used is often dictated by the facilities available, cost, environment (Windows, UNIX, etc.), and the quantities of data being managed. The software used to create and run SOMBASE comprises ORACLE 8i Enterprise, Microsoft Access 2000, Open Database Connectivity (ODBC) Drivers, and ArcGIS 8.2 (Oracle 2002, Microsoft 2000, ESRI 2002).

In the context of SOMBASE, Oracle is used as a relational database to perform the function of the primary data store. Any database format which is ODBC compatible can communicate with the GIS software. The primary data store for small databases could be a stand-alone PC running single-user software such as Access or Oracle Personal. Larger datasets that require a true multi-user facility, centralised data storage and back up copies will find using a server-held database more appropriate. SOMBASE is run in Oracle 8i Enterprise on a Sun Enterprise 250 machine with dual processors. It should be emphasised that this set-up is part of a multi-user facility within the British Antarctic Survey. Using this system it is possible to grant a range of levels of user privileges for different users from ‘selected tables read-only’ up to ‘full editing’ capabilities for added data security.

As this database is intended for use as a scientific resource it was decided that it was necessary to provide a user-friendly front end for entering, browsing and querying data. As most of the potential users of SOMBASE have a Windows based PC running the Microsoft Office suite of software, the obvious choice was MS Access 2000. Access provides a familiar user environment, with buttons and menus found in other MS Office software, as well as realtime connection to the Oracle datastore.

The connection between Access and Oracle is provided by ODBC Drivers. ODBC Drivers provide a communication layer between an application and a database server. The GIS component of SOMBASE is provided by ArcGIS 8.2 from Environmental Systems Research Institute (ESRI 2002). SOMBASE uses the full ArcInfo licence for ArcGIS. However, in most situations the less expensive ArcView licence will suffice. The GIS can be used for analysis, mapping and data editing. This advanced software is dependant on having hardware powerful and fast enough to run it, e.g. a minimum of a 450 MHz processor, a recommended 256 MB of RAM, and at least 16 MB of memory on the video card (ESRI 2001). Again ODBC drivers are used as an interface between the GIS and Oracle. Database queries, such as those created in Access, can also be connected to the GIS in the same way.

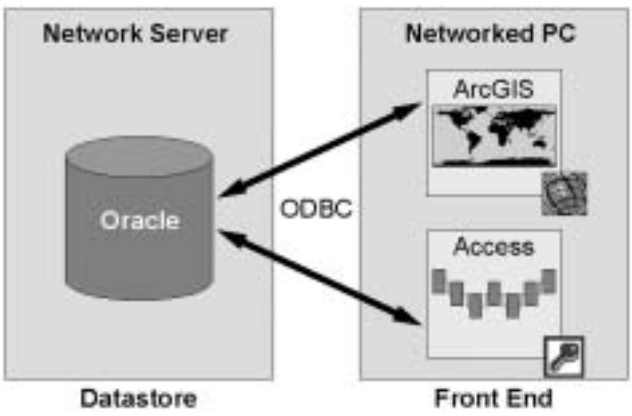

Fig. 2. Software and hardware layout of SOMBASE.

Applications

SOMBASE contains records for 950 species of Southern Ocean shelled gastropods, 136 species of bivalves and 2800 sites from different data sources such as original taxonomic, systematic and ecological papers, expedition reports, published and unpublished theses, and unpublished, archived raw data (see the SOMBASE reference list in Organisms Diversity and Evolution Electronic Supplement 03-12). The entered data can be viewed through the different Access tables (Fig. 3), higher-level tables, e.g. on the family level, can be browsed down to lower levels, e.g. the species level. On the species level all information on one species can be examined, such as synonyms (“other names”), shell size, shell characters,

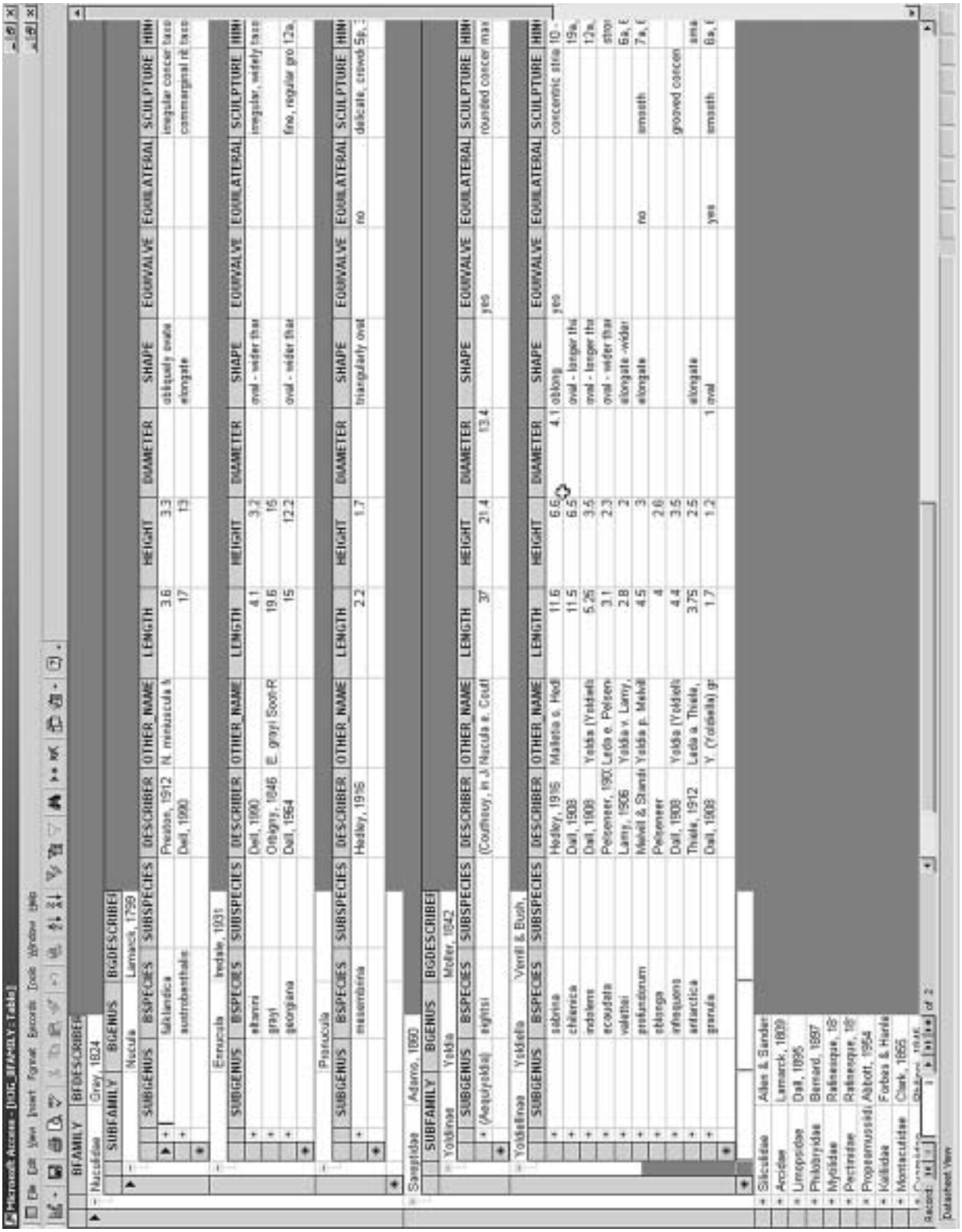

Fig. 3. Basic Microsoft Access tables for data entry and viewing.

and ecological features. These tables show the information gaps within the dataset and enable us to search specifically for the missing information.

The first step of analysis is to query the database on selected fields and then to create a report, for example in list form (Table 1). In our example the bivalve families Limopsidae and Philobryidae were queried on species depth and latitudinal ranges (Table 1). The report shows most of the limopsid species and the phylobryid genera *Adacnarca* and *Lissarca* to be eurybathic, while *Philobrya* is more restricted to shelf habitats. The database can be queried on each field included in the different tables (Fig. 1).

The second step is to analyse the data using GIS based queries to show distributions. Figure 4 shows all sites within SOMBASE at which bivalve or gastropod species have been recorded. The distribution records can be analysed using any suitable field of the key field tables. For example, the occurrence of selected taxa in predefined depth ranges can be displayed, or the occurrence of free spawners and direct developers along latitudinal transects. This type of analysis can uncover regional trends in functional groups.

An online directory of distribution maps for different systematic levels is available at www.antarctica.ac.uk/SOMBASE. The current system uses pre-generated images rather than a fully dynamic searching method creating maps on demand. The eventual aim is to have a website which allows the user to query the full database. Until this is possible the current selection of maps will be updated manually as more data is added to SOMBASE.

The use of a GIS enables the analysis and display of database queries in a more sophisticated way than those previously attempted. Selected areas can be studied by analysing chosen diversity patterns; in our example 5° latitudinal by 5° longitudinal area grids were analysed to

Table 1. Report resulting from a SOMBASE query on depth and latitudinal ranges for members of two selected bivalve families (Limopsidae & Philobryidae) occurring south of 60°S.

BFAMILY	BGENUS	BSPECIES	Min DEPTH	Max DEPTH	Min LAT	Max LAT
Limopsidae						
	Limopsis					
		hirtella	5	545	–65.95	–44.9667
		knudseni	2013	3693	–66.3833	–56.0667
		lilliei	110	535	–77.6567	–53.55
		mabilliana	842	1971	–53.0003	–31.1167
		marionensis	73	2804	–77.8333	–39.5333
		scabra	354	870	–77.0833	–72
		scotiana	101	342	–63.2889	–53.55
		tenella	2397	4328	–70.3333	13.1667
Philobryidae						
	Adacnarca					
		limopsoides	137	493	–78.4833	–61.2667
		nitens	9	2350	–78.4833	–53.7278
	Lissarca					
		aff. *miliaris*	494	494	–56.1	–56.1
		miliaris	9	4063	–64.8167	–52.5
		notorcadensis	5	1890	–77.7017	–53.5667
	Philobrya					
		atlantica	118	118	–54	–54
		capillata	126	463	–64.85	–53.7278
		cf. *kerguelensis*	199	199	–46.8067	–46.8067
		meridionalis	274	274	–51.31	–51.31
		multistriata	118	463	–54	–53.8583
		olstadi	75	75	–62.95	–62.95
		quadrata	16	267	–54.3667	–46.8067
		sublaevis	6	923	–78.3833	–53.7278
		wandelensis	110	110	–54.225	–54.225
	Verticipronus					
		tristanensis	12	140	–37.075	–37.075

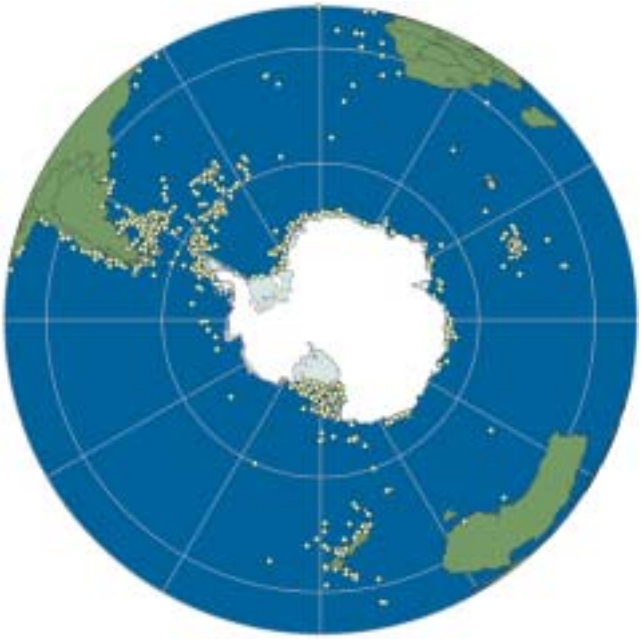

Fig. 4. Map of the Southern Ocean displaying the results of a geographical query. The dots represent sample locations of all bivalve and gastropod species included in the database.

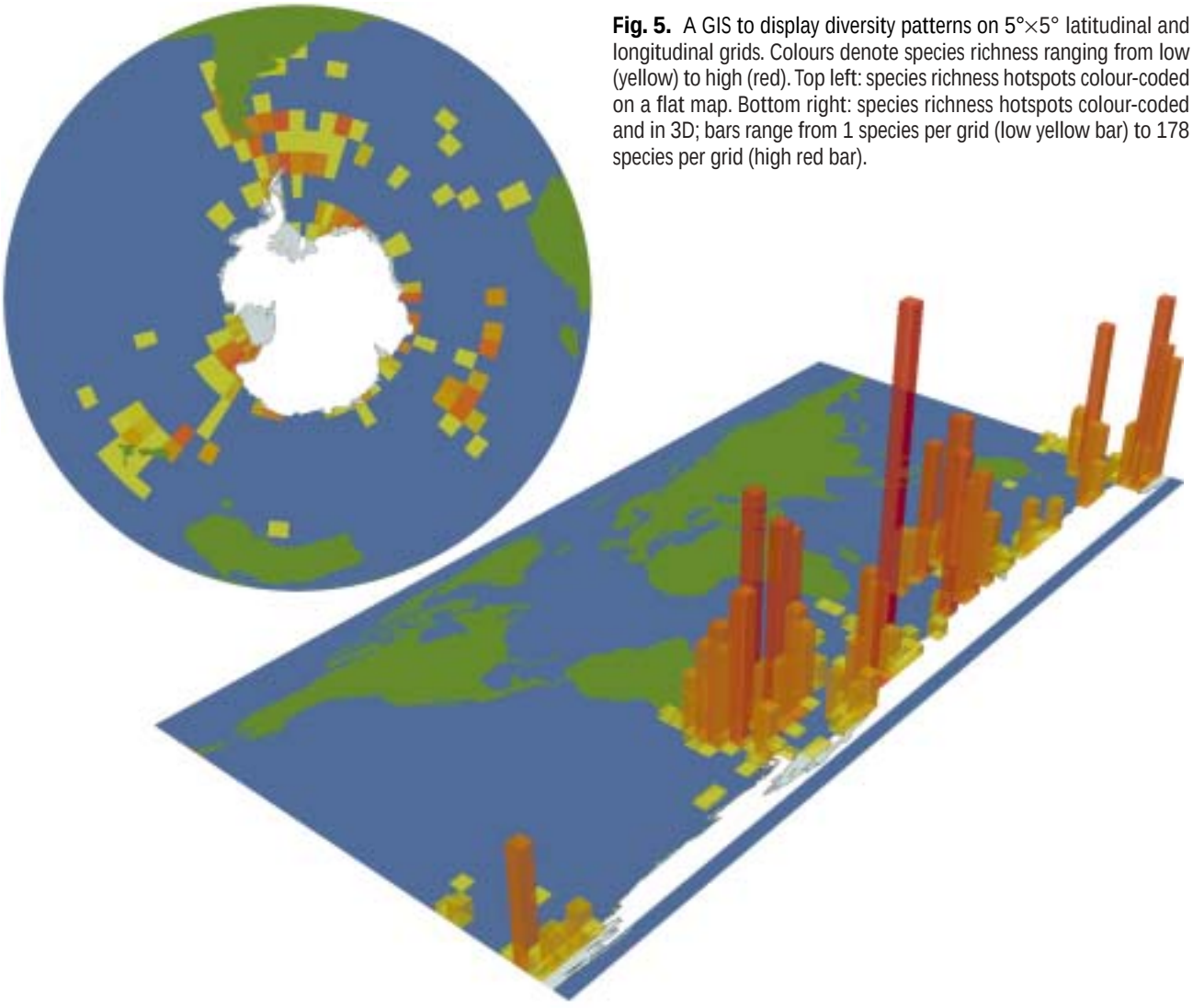

Fig. 5. A GIS to display diversity patterns on 5°×5° latitudinal and longitudinal grids. Colours denote species richness ranging from low (yellow) to high (red). Top left: species richness hotspots colour-coded on a flat map. Bottom right: species richness hotspots colour-coded and in 3D; bars range from 1 species per grid (low yellow bar) to 178 species per grid (high red bar).

identify hotspots of species richness (Fig. 5). Species richness, defined as number of species per selected area, can be displayed and colour-coded on a flat map (Fig. 5, top left), or in 3-D (Fig. 5, bottom right) to highlight differences between different hotspots. Diversity gradients can be visually pre-identified by comparing different area quadrats along a latitudinal transect or a longitudinal ring. To calculate if local species richness is a reflection of regional diversity the selected hotspot quadrats will be split into smaller area units and analysed in the same way.

SOMBASE incorporates the latest in database and GIS technology with a dataset spanning over a century. The new technology allows biogeographic data spanning vast areas to be visualised and analysed in an entirely new way. The initial results obtained from SOMBASE can be overlaid onto other datasets such as bathymetry, temperature, ice cover and primary production in an attempt to explain further the biogeographic patterns in the Southern Ocean. The only limitation on the potential analytical power of this tool is the availability and quality of the data. There is no reason why the same system cannot be expanded to cover even larger areas of the globe or different taxa. Conversely, it could be used for detailed studies of a single area or species.

References

Budd, A. F., Foster, C. T., Dawson, J. P.& Johnson, G. J. (2001): The Neogene marine biota of tropical America ("NMITA") database: accounting for biodiversity in paleontology. J. Paleont. 73: 743–751.

ESRI (2001): Performance Tips and Tricks for ARCGIS Desktop 8.1. http://www.esri.com

ESRI (2002): ARCGIS 8.x literature. http://www.esri.com

Grassle, J. F. & Stocks, K. I. (1999): A global ocean biogeographic information system (OBIS) for the census of marine life. Oceanography 12: 12–14.

Gutt, J., Sirenko, B. I., Arntz, W. E., Smirnov, I. S. & De Broyer, C. (2000): Biodiversity of the Weddell Sea: macrozoobenthic species (demersal fish included) sampled during the expedition ANT XIII/3 (EASIZ I) with RV 'Polarstern'. Ber. Polarforsch. 372.

Hill, M. J., Willms, W. D.& Aspinall, R. J. (2000): Distribution of range and cultivated grassland plants in southern Alberta. Plant Ecol. 147: 59–76.

Markwick, P. J. (2002): Integrating the present and past records of climate, biodiversity and biogeography: implications for palaeoecology and palaeoclimatology. In: Crame, J. A. & Owen, A. W. (eds) Palaeobiogeography and Biodiversity Change: a Comparison of the Ordovician and Mesozoic-Cenozoic Radiations. Geol. Soc. (London) Spec. Publ. 194: 179–199.

Markwick, P. J. & Lupia, R. (2002): Palaeontological databases for palaeobiogeography, palaeoecology and biodiversity: a question of scale. In: Crame, J. A. & Owen, A. W. (eds) Palaeobiogeography and Biodiversity Change: a Comparison of the Ordovician and Mesozoic-Cenozoic Radiations. Geol. Soc. (London) Spec. Publ. 194: 169–178.

Microsoft Corporation (2002): Microsoft Access 2000. http://www.microsoft.com/office

ORACLE (2002): ORACLE 8i Enterprise. http://www.oracle.com

Rosenberg, G. (1993): A database approach to studies of molluscan taxonomy, biogeography and diversity, with examples from Western Atlantic gastropods. Am. Malac. Bull. 10: 257–266.

Zhang, E. C., Shi, G. R., Sheng, S. & Zhou, W. (2000): Designing a computer-based information system for quantitative paleobiogeography of Permian brachiopods. Permophiles 36: 28–31.