THE DESIGN OF THE PMBC REFERENCE COLLECTION DATABASE

Somchai Bussarawit[1], Andrew Davison[2] & Charatsee Aungtonya[1]
[1]*Phuket Marine Biological Center (PMBC), P. 0. Box 60, Phuket 83000, Thailand, E-mail: pmbcnet@phuket. ksc. co. th.*
[2]*Dept. of Computer Engineering, Prince of Songkla University, Hat Yai, Songkhla 90112, Thailand*

ABSTRACT

The PMBC Reference Collection database holds over 14,000 records on marine species found in the south of Thailand. The design criteria for the database can be broadly classified into three groups: data issues, user interface issues, and system issues. Data issues include the database schema, extensions for new data types and multimedia support, the effects of the large dataset, integration with other database systems, the proposed search, update and viewing capabilities, the programmable features, and maintenance. User interface issues include the ease of use for people with differing technical backgrounds, on-line help, the availability of familiar windows-based input mechanisms. Output must be possible in a variety of forms, including as reports, tables and graphics. Planned LAN and Web access requires a consideration of security. System issues are concerned with the hardware and software support needed for the smooth running of the database. Hardware requirements include the machine configuration, backup facilities, peripherals, and network hardware. Software requirements involve the database runtime system, Multimedia, the operating system, and the network/Web interfaces. General system issues include reuse, stability, cost, speed, portability, and technical support. In this paper, we consider these various design issues in more detail, in order to motivate our choice of a custom-designed Access 97 database running under Windows 95 on a high-powered PC.

INTRODUCTION

The Phuket Marine Biological Center (PMBC) was established in 1968 in the south of Thailand under the auspices of the Thai Department of Fisheries, the Thai Ministry of Agriculture, and the Danish Government. PMBC holds the largest collection of marine animals in Thailand, and is a significant locus for taxonomic knowledge and biodiversity. Aside from its research into diverse areas of marine science, it also promotes the education of Thai and foreign scientists.

The number of specimens in the collection has rapidly increased, and now exceeds 100,000, distributed over 4,000 species. Specimen details were previously stored on registration cards, like the one shown in Figure 1. This physical form of data storage was sufficient in the early days of the center, but has become cumbersome and slow to use as the collection has grown. Also, it is hard to share the information on the cards due to the difficulty of duplicating and moving the information. Furthermore, its physical nature makes it very awkward to carry out complex searches over the data or perform analyses upon it. Printing specimen catalogues extracted from the cards is also problematic. For these reasons, we decided to convert the antiquated card system into a database form. Our long-range plan (ten years and beyond) is to store up to 40,000 species in the database, and to make it accessible via CD-ROM and the Internet.

We have built a custom-designed relational database using Microsoft Access 97 (Gifford

REGISTRATION CARD PMBC _____

FOR THE REFERENCE COLLECTION

Phylum	Institution Donated by	Coll. Number Ph-Number Photograph Number
Class	Order	Family
Species		
Locality		
Collector	Date	Notes
Identifier	Date	Notes
Confirmed by	Date	Notes

Figure 1. A PMBC Reference Collection Card (The back of the card contains space for field data notes.)

et al. 1997; Jennings 1995) running under Windows 95. Most of the search capabilities and the user interface have been completed, and details from over 14,000 specimen cards have been entered into the database in the six months since the project's inception. In this paper, we describe the criteria behind our design in terms of three main areas: data issues, user interface issues, and system issues.

DATA ISSUES

The choice of data format (or in database parlance, the *schema)* was the first important decision, and was based on reproducing the fields in the existing registration card system. The key index (or *primary key)* of a card was the PMBC number, and this has been retained since there is no other obvious alternative candidate. During a preliminary state, the specific part of the binomem was considered (Jalk and Nateewathana 1995), but unfortunately it is conceivable that this may change on a card over time (due to a reclassification perhaps). The only constants for a card are the PMBC number, and minor fields such as the identification date and the field notes. Of these, the PMBC number is the simplest primary key. An interesting practical observation is that the speed of Access makes the choice of primary key less essential than will other packages - its speed when searching, using other fields (such as identification data or taxonomic classification) is very fast.

Another decision was whether to subdivide the database into several parts, but after careful consideration this did not seem necessary: the principle unit of information in the database is the 'card' for a specimen, and there is little advantage in subdividing this into multiple objects. One issue was whether to separate out specimen-independent taxonomic information, but it proved more useful to retain it with the rest of the specimen data, The main reason is that if a specimen is wrongly identified then the taxonomic information need only be changed on the card itself. A description of database design methods, and issues such as normatisation, can be found in (Date 1995).

Another reason for retaining the card format

is its inherent good design - most specimen collections require information like that used on the PMBC reference card, and it would be straight forward to use this data design for storing, other types of specimens, not just marine animals.

The database format was extended in several ways, mostly with extra fields to reflect the changes in the specimen data as it is maintained over time. For example, additional fields were included to allow for second identifiers, and for a reclassification of the species name. It was felt that the initial species identification should be retained, since it can be useful for certain kinds of data analyses.

An important reason for choosing Access 97 is its support for 32-bit OLE. OLE (Object Linking and Embedding) is the Windows mechanism for transferring and accessing information between different applications, and allows Access to store multimedia data in its databases. A planned extension of the database is to store a picture for each specimen, and perhaps video and sound.

Access was also chosen because of its powerful data import and export features. Data import means that Access can load data stored in most database styles (e.g. dbase, Paradox, Btrieve, plain text) and convert it to Access database form with only minor user intervention. We used this feature to import some data from an old DOS-based dbase III database. Access can also export data in a variety of formats, including as Microsoft Word files and Excel databases. The database uses the Word export capability when generating reports.

There are several existing taxonomic packages, such as Platypus (CSIRO 1995) and Linnaeus 11 (Estep *et* a/. 1993). They offer a great deal of general power and flexibility, but lack specific features required by researchers within PMBC. One example is the need to count the number of taxonomic groups (the number of species, families, and so on) which meet userspecified criteria. This lack of focused aspects was a significant motivation for building, a new database package, fine-tuned for the needs of PMBC. This is most apparent in the types of search available to the user.

Searches by taxonomy can be specified in terms of phylum, class, order, family or species, and a list of matching, cards are generated. The full details for a particular card are only a button press away, as is the ability to print the matching information or save it in Word format. The printed output is formatted in the familiar taxonomic hierarchy of phylum, class, order, family and species used in catalogues, the idea being to make it simple to use the database output as the basis of publications. The lists of taxonomic categories available in the searches are generated at search-time, and so always represent the current knowledge of the database.

There are several searches based on dates, such as searches for cards added before, on, or after a certain collection date or identification date. Searches between date ranges are also possible. As with the taxonomic searches, a summary of matching cards is presented, which can be readily examined in more detail. Word fragment searches look for the specified text within the notes and field data fields. Place searches look through the locality and province fields of every card. Count searches return counts and lists of matching cards based on user criteria. The lists are for the taxonomic groups (e.g. the distinct families) which match the criteria.

These search utilities satisfy the needs of most novice users (although user evaluation is still being carried out at the time of writing). For more advanced users, Access's 'filter by form' and 'filter by selection' mechanisms are available. These are quite hard to use in their full generality, and so their interface in our database is based on the familiar old-style 'card' layout (see Fig. 2). With these filtering, methods, the user can enter search criteria into any of the fields of the 'card' shown on the screen in order to

filter down the total number of results. This process can be repeated until the filtering produces the matches of interest. Powerful (but obscure) search techniques using wild cards and boolean operations are possible, and the aim of the 'card' screen layout is to make them more intuitive to use. On-line help is supplied for these features.

The database's search capacity utilises Access's extensive database programming capabilities, including SQL queries, macros and Visual Basic for Applications. This gives unprecedented control over all features of the database, and even other applications on the machine. The database can be updated in several ways: there is a 'card'-based interface for the entry of information by novice users, but more advanced users can edit the database in table form. Care has been taken to separate the search and update features, so that data can be protected. This separation will become more important when the database is networked and connected to the Web.

USER INTERFACE ISSUES

The ease of use of the database is of primary concern, since it is aimed at people with a detailed knowledge of marine science but little familiarity with databases. Consequently, the interface places heavy emphasis on the use of menus, buttons, and on-line help, combined with a look-and-feel closely aligned to the old 'card' system. The on-line help is context sensitive, and always only a button press away. It is presented in the familiar Windows 95 format, complete with detailed cross-referencing and examples.

A long-term aim is to make the database available across the PMBC LAN, and ultimately accessible to anyone via the World Wide Web. This latter aim highlights PMBC's

wish to promote knowledge of South-East Asia marine biodiversity.

Access was chosen because of its excellent support for networking (e.g. features such as database replication, sharing, splitting, encryption, passwords). It can readily employ the built-in networking of Windows 95 (or Windows NT). This has not yet been carried out, but the database has been designed with this aim in mind. It is also hoped to make the database available on a CD-ROM, and this need not require other users to possess a copy of Access; a simplified run-time version is available for a nominal fee, or a stand-alone database application can be created. Access 97 possesses excellent Web support.

SYSTEM ISSUES

The choice of Access means that the hardware and software requirements for the database become reasonable. We utilise a standard COMPAQ PC with a Pentium 11 chip and 40 MB of RAM. This results in an extremely fast database, where even complex searches take at most 2-3 seconds. The planned multimedia needs of the database require speakers and good screen resolution, but these are standard accessories in today's PCs. We use a SyQuest external disk drive for backups, a scanner and digital camera for inputting photographs, and a laser printer for output. Software costs on a Windows platform are generally lower than for specialised workstations and their operating systems. Access can utilise other Windows software quite simply via its support for OLE. Access also comes with some useful analysis tools including utilities for detecting performance bottlenecks, carrying out database compaction, and for splitting a database prior to networking. Several other tools are included in the Access developer's toolkit, including the Access runtime system, a help compiler, a package for creating stand-alone applications, and a replication manager. One disappointment was the overly complex help compiler, and we eventually moved to a shareware product called HelpScribble (Goyvaerts 1997) which is considerably easier to use.

Many of our choices were based on pragmatic concerns such as the wish to use hardware and software already familiar to users at PMBC and elsewhere. Access is a good choice even for non-database users because of its similar interface to other Microsoft products such as Word and Excel. The stability of the hardware and software was a concern due to the costs of technical assistance, and the lack of expert local help. Windows 95 is currently on its third minor upgrade and is much more robust than when it was first released. The choice of Windows means that some local technical support is available.

ACKNOWLEDGEMENTS

We wish to thank PMBC and DANIDA for supporting this database project.

REFERENCES

CSIRO. 1995. "Platypus: A Database Package for Taxonomists", *CSIRO Publishing,* Australia. See http://www.ento.Csiro.au/platypus/platypus.html

Date, C.J. 1995. *An Introduction to Database Systems,* (6th edition), Addison-Wesley.

Estep, K.W., R. Sluys, & E.E. Syvertsen, 1993, "Linnaeus and Beyond: Workshop Report on Multimedia Tools for the Identification and Database Storage of Biodiversity". *Hydrobiologia,* 269/270: 519-525.

Gifford, D., *et al.* 1997. *Access 97 Unleashed,* SAMS Publishing, USA. 1082 pp.

Goyvaerts, J. 1997. "HelpScribble: a help authoring tool, which works together with the help compilers from Microsoft." See http://www.jgsoft.com/helpscr.shtml

Jalk, H.R. & A. Nateewathana, 1995. Database of a Reference Collection — The Advantage of Having Specific Names as Main Entry. Phuket mar. biol. Cent. Spec. Publ. **15**: 89-92.

Jennings, R., 1995. *Special Edition Using Access **95**,* Que Corporation, USA.