A microcomputer-based relational database for an academic department of human genetics

F. VAN GREUNEN, K. J. MACGREGOR, P. BEIGHTON

Summary
Recent advances in microcomputer technology have made it possible for academic departments to establish their own discrete databases. This is in keeping with the modern tendency towards distributed processing on smaller systems as opposed to depending on large shared remote centralized mainframes. A database that had been implemented on a mainframe for the Department of Human Genetics of the University of Cape Town has been successfully transferred to a microcomputer system, resulting in a redesigned relational system with several significant advantages. These include faster data capture, enhanced consistency, greater computer awareness, improved economy and increased confidentiality.

A record system for the Department of Human Genetics of the University of Cape Town (UCT) had been designed and implemented using the fourth-generation language MAPPER on the Sperry mainframe computer of UCT. Recently microcomputer technology has shown rapid development and good database packages specifically designed for microcomputers have become available.

Some practical aspects and difficulties that became apparent during use of the original implementation, as well as the feasibility of transferring the database to a microcomputer system in an attempt to alleviate these problems, are discussed. The design features of a microcomputer-based database management system (DBMS) that has evolved in the Department of Human Genetics are described.

Practical aspects — MAPPER v. microcomputer

The following undesirable features became apparent with the original implementation under MAPPER for which a microcomputer might provide solutions.

Limited record types. MAPPER limits the user to a mere 8 record types (FORMS in MAPPER data definition language) in a particular partition (MODE in MAPPER data definition language).

Limited record size. MAPPER limits the user to a record size of 152 characters which was found to be inadequate for our purposes.

User interface. The data manipulation language (DML) of MAPPER is difficult for casual users of the system.

Teleprocessing costs. Communication costs necessitated by the hire of modems and dedicated telephone lines are high. Similarly the ultimate cost of using secondary storage media on the mainframe is not insignificant.

Confidentiality of sensitive patient data. In view of the highly sensitive nature of the personal patient data and in the light of the international concern, it was felt that the information should preferably be maintained on an in-house database.

Lack of other computing facilities. MAPPER limits usage to database administration unless use is made of the privileged 'Remote Symbiont Interface' facility. This is regarded as inequitable with respect to other users.

Data independence. Alteration to the conceptual schema of the database with concomitant maintenance of data independence is difficult with MAPPER. Even minor changes to the conceptual schema call for time-consuming planning and meetings with the MAPPER co-ordinator.

Availability. MAPPER on the UCT mainframe is available at certain times only, precluding database management after hours and over weekends and on holidays.

Implementation on a microcomputer

Some terms from elementary relational database theory will be used in this article but will not be defined since they are adequately described in standard works.

Hardware

An IBM PC microcomputer system consisting of the central processing unit with 192 kilobytes of random access memory and two floppy disk drives each with a capacity of 360 kilobytes and operating under the PC-DOS 2.0 operating system was available on a limited scale. The system is used extensively for word-processing in the Department. It is planned to change to a system with a quick access fixed disk (20 megabyte) in the near future.

Software

The relational database package dBASEII was chosen because it has proved to be pre-eminent in the small-computer field. In addition, although the DML is somewhat cumbersome, a powerful programming facility exists with which more 'user-friendly' protocols can be written for retrieval, external view creation and other database functions. As with any DBMS, however, these functions make the existence of a database administrator responsible for such duties highly desirable.

Adequate space is provided by the package. The number of different relations that can be created is virtually unlimited, records (tuples) can contain up to 1000 characters and there is provision for up to 32 fields (attributes) per record. These specifications were considered to be adequate for our requirements.

The application programs were written using the programming facilities offered by dBASEII and details will be published elsewhere.

Department of Human Genetics, University of Cape Town
P. BEIGHTON, M.D., PH.D., F.R.C.P., D.C.H.
Department of Computer Science, University of Cape Town
K. J. MACGREGOR, M.SC.
Design and lay-out of the database

Original design

Table I shows the structure of the database as originally conceived and as produced by the display structure command of the DML. Apart from some minor differences, this is the same as that which had been used on the mainframe under MAPPER. With dBASEII, however, all the attributes could be accommodated by one relation (table) whereas under MAPPER the constraint of 132 characters per record required that the data be spread over two report identifiers. The meanings of those attribute names not immediately apparent are given as comments after asterisks in Table I.

| TABLE I. COMPLETE CONCEPTUAL SCHEMA |
|---|---|---|
| FLD | NAME | TYPE | WIDTH |
| 001 | ID | N | 006 |
| 002 | NAME | C | 020 |
| 003 | BIRTHDAY | N | 006 |
| 004 | SEX | C | 001 |
| 005 | ETHGROUP | C | 001 |
| 006 | SUBGROUP | C | 010 |
| 007 | CENTRE | C | 037 |
| 008 | CLINIC | C | 010 |
| 009 | FIRSTDATE | N | 006 |
| 010 | EXAMNER | C | 002 |
| 011 | REASON | C | 002 |
| 012 | GROUP | C | 010 |
| 013 | GROUPCODE | C | 002 |
| 014 | SPACER | C | 001 |
| 015 | DIAGCODE | N | 005 |
| 016 | DISCONFIRM | L | 001 |
| 017 | FOLDER | C | 001 |
| 018 | LASTDATE | N | 006 |
| 019 | DEATHDATE | N | 006 |
| 020 | PATTERN | C | 001 |
| 021 | CONSA | C | 001 |
| 022 | KB | C | 001 |
| 023 | ANTENTAL | L | 001 |
| 024 | KEYS | C | 001 |
| 025 | CHROMES | C | 001 |
| 026 | ME | L | 001 |
| 027 | PH | L | 001 |
| 028 | SL | L | 001 |
| 029 | PU | L | 001 |

E = CHARACTERS N = NUMERICAL L = LOGICAL

Table II shows the same relation with some of the attributes removed since these are irrelevant to the rest of this article and facilitate the presentation of database views as well as the reproduction of the computer print-outs.

The information visible to a casual browser through the database is presented in Table III. We use the word 'browser' since 'browse' is a verb in the dBASEII DML which is very easily implemented by simply typing the command at the terminal. It is clear that confidential information was not adequately protected since it was easy for anyone who could access the database to 'peregrinate' through the database at will with names and diagnoses in full view.

| TABLE II. ABRIDGED CONCEPTUAL SCHEMA |
|---|---|---|
| FLD | NAME | TYPE | WIDTH |
| 001 | ID | N | 006 |
| 002 | NAME | C | 020 |
| 003 | BIRTHDAY | N | 006 |
| 004 | SEX | C | 001 |
| 005 | ETHGROUP | C | 001 |
| 006 | SUBGROUP | C | 010 |
| 007 | CENTRE | C | 037 |
| 008 | CLINIC | C | 010 |
| 009 | FIRSTDATE | N | 006 |
| 010 | EXAMNER | C | 002 |
| 011 | REASON | C | 002 |
| 012 | GROUP | C | 010 |
| 013 | DIAGCODE | N | 005 |
| 014 | DISCONFIRM | L | 001 |

C = CHARACTERS N = NUMERICAL L = LOGICAL

New design

Maintaining confidentiality

The problem of confidentiality was approached by the use of passwords and by the application of relational database principles. Access to the database is obtained only after entering a password for which a compiled program prompts. While someone with a fair knowledge of the operating system can circumvent the password, it is not anticipated that there would ever be enough motivation to do so.

While Table I still represents our conceptual schema, i.e. what is available on the database, we have restructured the database into four relations:

Table IV shows the structure of the main relation named 'mainbase' (abridged as explained above) which is that of Table II but with names, specific diagnoses and group diagnoses removed. Table V shows the information which the casual browser will encounter having accessed the mainbase. It is clear that patients' names and diagnoses are more adequately protected than was the case on the original implementation and on MAPPER.

Table VI shows the structure of a new relation named 'diagnoses' with the two attributes diagnosis 'code' and 'diagnosis'. The data in this relation is based on McCusick's classification with a
VI. RELATION DIAGNOSES

<table>
<thead>
<tr>
<th>CODE</th>
<th>DIAGNOSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>25278</td>
<td>MPS I (HURLER SYN)</td>
</tr>
<tr>
<td>25290</td>
<td>MPS III (SAN FILIPPO SYN)</td>
</tr>
<tr>
<td>25300</td>
<td>MPS IV (MORQUIO SYN)</td>
</tr>
<tr>
<td>25320</td>
<td>MPS VI (MARTEAUX-ARYM SYN)</td>
</tr>
<tr>
<td>25360</td>
<td>LIMB GIRDLE MUSCULAR DYSTROPHY</td>
</tr>
<tr>
<td>25580</td>
<td>SCHWARTZ SYN</td>
</tr>
<tr>
<td>25720</td>
<td>NIEMANN-PICK DIS</td>
</tr>
<tr>
<td>25770</td>
<td>GOLDENHAR SYN</td>
</tr>
<tr>
<td>25920</td>
<td>BLOUNT DIS</td>
</tr>
<tr>
<td>25970</td>
<td>OP AR (OSTEOPETROSIS ALBERS-SONBERG DIS AR)</td>
</tr>
<tr>
<td>26160</td>
<td>PKU (PHENYLKETONURIA)</td>
</tr>
<tr>
<td>26340</td>
<td>THALASSEMIA</td>
</tr>
<tr>
<td>26352</td>
<td>SHORT RIB POLYDACTYL SYM</td>
</tr>
<tr>
<td>26440</td>
<td>PSEUDOHYPOPARATHYROIDISM</td>
</tr>
<tr>
<td>26480</td>
<td>PXE AR (PSEUDOXANTHOMA ELASTICUM AR)</td>
</tr>
<tr>
<td>27005</td>
<td>RUSSEL-SILVER SYM</td>
</tr>
<tr>
<td>27020</td>
<td>SJOGREN-LARSSON SYM</td>
</tr>
</tbody>
</table>

TABLE VII. RELATION NAMES (FICTITIOUS DATA)

<table>
<thead>
<tr>
<th>ID</th>
<th>NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1800</td>
<td>ADAMS PETER</td>
</tr>
<tr>
<td>1801</td>
<td>VAN WYK JAN</td>
</tr>
<tr>
<td>1802</td>
<td>SAAYMAN ANNA</td>
</tr>
<tr>
<td>1803</td>
<td>PATEL PRIYNEHA</td>
</tr>
<tr>
<td>1804</td>
<td>NEL GERHARDUS</td>
</tr>
<tr>
<td>1805</td>
<td>JONES JOHN</td>
</tr>
<tr>
<td>1806</td>
<td>SMITH KATE</td>
</tr>
<tr>
<td>1807</td>
<td>NEL PETRUS JOHANNES</td>
</tr>
<tr>
<td>1808</td>
<td>VAUGHAN SARAH</td>
</tr>
<tr>
<td>1809</td>
<td>BOOLEY ZAIDA</td>
</tr>
<tr>
<td>1810</td>
<td>BARNARD KATRINA</td>
</tr>
<tr>
<td>1811</td>
<td>RUITERS SUSAN</td>
</tr>
<tr>
<td>1812</td>
<td>STRYDOM GERHARD</td>
</tr>
<tr>
<td>1813</td>
<td>MOLLEF AMOS</td>
</tr>
<tr>
<td>1814</td>
<td>RUITERS DAPHNE</td>
</tr>
</tbody>
</table>

locally prepared extension for non-genetic disorders. Given the diagnosis code, which acts as primary key, the diagnosis can be looked up by the computer and, conversely, an attempt may be made to get the diagnosis code, given the diagnosis. We say 'attempt' because with the converse, spelling of the diagnosis becomes critical and success cannot be guaranteed.

The relation 'diagnoses' was created by letting the data capturers enter the patients' diagnoses for the first 500 cases. A program was then written which extracted a nucleus of diagnoses and matching codes into the new relation. This nucleus was subsequently carefully edited. At present, with a nucleus of diagnoses and codes at hand, only the codes are entered for new patients. If a code is encountered which is not present in the relation, the data capture program will prompt for the name of the diagnosis and then append a new record to the relation.

The unabridged database has an attribute 'group diagnosis' to which the same considerations apply. A relation similar to 'diagnoses' has been implemented for the group diagnoses and will not be detailed.

Serendipity

Using the previous two relations has proved to have two unforeseen advantages: input effort is reduced as the data capturers need not enter specific and group diagnoses for the patients. In addition consistency is considerably enhanced since the use of a code rather than a name to a large extent dispels the difficulties in retrieval caused by spelling and abbreviation inconsistencies. These findings are in keeping with a high level of database normalization which is known to reduce redundancy and improve consistency. Normalization to at least third normal form has been obtained if we regard ID and DIAGCODE as a composite primary key.

Table VII shows the structure of the fourth relation named 'names' with two attributes namely ID (primary key) and NAME. The access path to this relation is password-controlled so that only privileged users are given access to patients' names.

Query examples

Two examples of database queries are given in crude algorithmic form. For brevity the customary steps that deal with error conditions are omitted and it is also assumed that all relevant records exist.

1. List IDs of all cases with a certain specific diagnosis: Table VIII shows the algorithm where the diagnosis is 'PXE' (pseudoxanthoma elasticum). Note that, should the search have been direct on MAINBASE using 'PXE' as search string, the case with ID 1807 would have been missed because of a spelling inconsistency (see Table III).

   Table VII shows the structure of the fourth relation named 'names' with two attributes namely ID (primary key) and NAME. The access path to this relation is password-controlled so that only privileged users are given access to patients' names.

   The unabridged database has an attribute 'group diagnosis' to which the same considerations apply. A relation similar to 'diagnoses' has been implemented for the group diagnoses and will not be detailed.

   Serendipity

   Using the previous two relations has proved to have two unforeseen advantages: input effort is reduced as the data capturers need not enter specific and group diagnoses for the patients. In addition consistency is considerably enhanced since the use of a code rather than a name to a large extent dispels the difficulties in retrieval caused by spelling and abbreviation inconsistencies. These findings are in keeping with a high level of database normalization which is known to reduce redundancy and improve consistency. Normalization to at least third normal form has been obtained if we regard ID and DIAGCODE as a composite primary key.

Table VII shows the structure of the fourth relation named 'names' with two attributes namely ID (primary key) and NAME. The access path to this relation is password-controlled so that only privileged users are given access to patients' names.

Query examples

Two examples of database queries are given in crude algorithmic form. For brevity the customary steps that deal with error conditions are omitted and it is also assumed that all relevant records exist.

1. List IDs of all cases with a certain specific diagnosis: Table VIII shows the algorithm where the diagnosis is 'PXE' (pseudoxanthoma elasticum). Note that, should the search have been direct on MAINBASE using 'PXE' as search string, the case with ID 1807 would have been missed because of a spelling inconsistency (see Table III).

   Table VIII. Algorithm

   1. Enter diagnosis on computer prompt;
   2. Change to capital letters (*pxe* may have been entered *);
   3. Access relation DIAGNOSES;
   4. Locate record where diagnosis field contains 'PXE';
   5. Store the corresponding code in a search variable;
   6. Access relation MAINBASE;
   7. Traverse MAINBASE, printing each ID where DIAGCODE matches the search variables;
   8. Exit.

2. List names of patients seen by a certain examiner: Table IX illustrates the algorithm.

   Similar as well as more complex views are readily created by using the dBASEII programming facility. Some examples of such queries are:

   1. List the IDs of all white patients with 'PXE' or 'EDS' (Ehlers-Danlos syndrome) examined by 'PB' or 'FVG' at 'GSH' (Groote Schuur Hospital). The abbreviations are unique identifiers as agreed upon by in-house convention.
   2. List the names of female patients of all ethnic groups with dwarfism born between 1954 and 1976.

Conclusion

At the time of writing about 1500 cases have been entered without major difficulty. Performance of the microcomputer and database is extremely satisfactory. The IBM floppy disks each accommodate 1000 cases together with appropriate indexes and control files with ease. This is adequate but may become
TABLE IX. ALGORITHM 2.

List the names of patients seen by a particular examiner.

1. Enter password on computer prompt ("patients’ names are involved necessitating password control");
2. If password accepted, go to step 3, otherwise exit;
3. Enter examiner’s initials on computer prompt and change to capital letters if necessary;
4. Access relations MAINBASE and NAMES;
5. Do steps 6 to 9 until end of MAINBASE is reached;
6. Locate the next record in MAINBASE with examiner field matching the initials;
7. Store the ID of the record in a search variable;
8. Locate the record in NAMES where ID matches the search variable;
9. Print the NAME field of the record;
10. Exit.

3. Increased computer awareness and enthusiasm for electronic record keeping have resulted.
4. Accuracy of data is very much enhanced as frequent printouts in a variety of permutations can be produced and subsequently critically reviewed and edited.

The small computer is here to stay and rapid development in speed and power of application is foreseen. Renewed interest has been kindled for relational database theory and application as originally conceived by Codd and others, and rapid strides are expected in this field. The small system is in keeping with the modern tendency towards distributed processing as well as with Tofflerian third-wave philosophy. The concept of deferred design on which computer development was based has resulted in a domain of possible applications which is delightfully protean, the benefits of which can only be realized when a system, albeit small, is physically present in a department.

This project was supported by the South African Medical Research Council, the Mauerberger Foundation Fund, the Harry Crossley Foundation Fund and the University of Cape Town Staff Research Fund.

REFERENCES