

Institute of Applied Health Sciences

University of Aberdeen

DATABASE REVIEW

Institute of Applied

Health Sciences

University of Aberdeen

DATABASE DESIGN

GUIDELINES

Contents

| | |
|--------------------------|-----------|
| Introduction | 3 |
| Overview | 3 |
| Ethical Approval | 3 |
| Basic Considerations | 3 |
| Confidentiality | 5 |
| Data Protection Act 1998 | 5 |
| Data Security | 5 |
| Database Function | 6 |
| Input | 6 |
| Objectives | 6 |
| Output | 7 |
| Database Design | 8 |
| Normalisation | 8 |
| Data Quality | 8 |
| Performance | 9 |
| Documentation | 11 |
| Database Testing | 12 |

Introduction

Overview

Database design theory is a fairly dry subject, but a basic understanding is necessary to avoid some of the pitfalls encountered when a new database is created. This document provides strategic guidelines for the basic design and development of databases within the Institute of Applied Health Sciences.

Further advice on functional analysis and database design, for existing databases or for future planned databases, can be obtained from the Data Manager for the Institute of Applied Health Sciences (Val Angus, ext 59772, Email: v.angus@abdn.ac.uk).

Ethical Approval

Ethical Approval and Funding should be sought well in advance of the project, although a clear idea of the envisaged database function and a description of the data items to be collected should be available.

The **Confidentiality Guidelines** document issued by the Institute of Applied Health Sciences should be consulted prior to obtaining and recording confidential information on paper or on a computer application package. Confidentiality should be the prime concern for the research project as failure in this aspect could jeopardise the credibility of the Institute and may lead the University of Aberdeen to breach legal requirements, with a consequent impact on future research.

Basic Considerations

Sound database design is crucial to the realisation of the potential benefits of the recorded information. It is important for database designers to have conceptual clarity, without this databases may evolve into tangled webs where information retrieval becomes a nightmare. The database designer should always keep in mind the kind of information retrieval that will be required and ensure that data is held in a suitable structure to facilitate envisaged queries and any possible future queries.

One way of minimising the possibility of tangled structures, is to ensure that data tables are normalised removing repeating groups and redundancy. This procedure provides data stability as the database grows, changes and is used for new applications. In a fully normalised database, specific data items appear in one place only allowing efficient update and consistency of data.

Data quality is essential for accurate information. In addition to referential integrity, validation of information entered on the database is vital and should constrain data input to specific ranges or values, field sizes and type.

Consideration should be given to the location of the new database, ranging from the hard drive of a personal PC to a secure central server. The latter option has definite advantages with regard to security, data linkage, ease of maintenance, automated back-up facilities and data recovery.

An estimate of database size may be determined from the number of expected records multiplied by the number of storage bytes per record. An estimate of the yearly growth rate should also be included, together with a contingency factor of two to ensure that there is sufficient disc space available for the estimated size of the mature database.

Institute of Applied Health Sciences

University of Aberdeen

Most smaller databases, less than 30 Mbytes, are created in Microsoft Access or, alternatively, in Statistical Package for Social Scientists (SPSS) for questionnaire oriented research projects. Microsoft Access can be used for databases up to a maximum of 1 Gbyte, but database performance is degraded as the size of the database increases. SQL Server, Oracle or SIR should be considered for databases expected to grow larger than 30 Mbytes, with SQL Server being the preferred option due to its ease of linkage with other applications, access via the internet, user-friendly software and widely available expertise.

Confidentiality

Confidentiality should be the prime concern for the research project as failure in this aspect could jeopardise the credibility of the Institute and may lead the University of Aberdeen to breach legal requirements, with a consequent impact on future research.

Data Protection Act 1998

The Data Protection Act 1998 legislated specifically on the treatment and use of confidential data held in a structured way in any medium (paper, computer, microfiche, tape etc). The implications of this legislation on clinical research is summarised in the “**Confidentiality Guidelines**” document issued by the Institute of Applied Health Sciences.

All staff and students involved in a research study with potential access to confidential information must sign a Confidentiality Statement and similarly, the database custodian must complete and sign a Confidentiality Protection Statement for the database to be created. The originals of these Statements must be returned to the Institute Co-ordinator, with copies retained by the individuals.

Informed written consent must be obtained for all subjects participating in a research study before information is recorded on a database.

Data Security

Databases containing confidential information must be password protected to prevent unauthorised access, with every user having his or her own database account. Similarly, where network access is to be utilised, the network should be protected from external unauthorised access.

Isolate non confidential information into separate tables or databases which could potentially be shared with other users, accessed from a central source. For example, Clinical coding systems are common to clinical databases and tables could be held centrally and made available for other users, reducing duplication of maintenance for the common data tables. Sharing information resources and non confidential data saves time and money.

Data must be protected from being modified by individuals who have no right to modify them. Read only and Read / Write access can be configured via passwords. Additional security authorisations can be applied to confidential tables allowing restriction of different levels of users to selected views of the whole database.

Frequency of password changes should be determined, ideally with the database system prompting for password changes at 30 day intervals.

Procedures must be in place to recover data in the event of accidental loss or damage. Routine daily security back-ups are recommended and for larger databases, daily file activity logs should be kept to avoid the loss of a complete day's transactions.

Database Function

A common problem with the design of databases is that they are created to record data and no consideration is given to subsequent data retrieval. Consequently, after a period of time, the researcher may find that he/she has recorded a lot of valuable information but cannot realise its potential without considerable programming support to disentangle the data items.

It is important to define the required function of the database at an early stage and to list the kind of information retrieval that is envisaged to ensure ease of access at a later date.

The possibility of accessing non-confidential reference data from a central source should be explored. For example Clinical Classifications, Standard Occupational Classifications, Deprived Area postcodes and GP Practice addresses.

Input

It is important to review all the possible sources of input to the new database, ranging from paper documents/forms to electronic interfaces with external systems. The interface between the database and the user must be “user-friendly” and allow efficient input with suitable validation and domain checks to ensure best possible data quality. Provision of “drop down” boxes for data entry will improve speed and consistency of data entry.

Particular attention should be paid to the feasibility of assigning default values, automatic tab control and automatic capitalisation of initial letters and postcodes.

Referential integrity between tables should be carefully planned to ensure that data is recorded on the database in a consistent form. Where data entry is inhibited due to referential integrity, the user should be given the opportunity to input the missing data in the referenced table.

Objectives

The overall function of the database and data flows should be clarified at the outset to ensure that the initial database design encompasses all of the envisaged requirements.

A simple menu structure covering all of the main functions of the database is a desirable feature. Where codes are used to store information, the full description corresponding to the code should be visible on data entry and output displays.

There may be a requirement to use Mail Merge for contacting patients and if the database is constructed with a single compound address field, it is inefficient to have to separate the data into 4 individual address lines and a postcode field prior to export to a Mail Merge package.

Some information to be recorded, such as drug regimes, problems or surgical operations, may be time-related, so where possible, the database should include temporal information to facilitate meaningful queries or allow data comparisons for longitudinal studies.

Use of indexes should be evaluated to increase speed of information retrieval, but this must be balanced with degradation on write performance.

Output

Output is one of the most important considerations in database design. Where possible, a standard user interface should be designed to allow generation of standard reports. Codes used as foreign keys to minimise data storage should be expanded to the full description on user reports and display screens.

Database tables should be designed for ease of information retrieval, with data structured in a consistent unambiguous form.

Where possible, simple monitoring systems should be implemented to highlight exceptions or to flag patients outwith the normal range of values for the clinical research project. Immediate feedback of this kind can be very helpful in clinical research.

Routine calculation of statistical trends and distribution of variables can be a useful source of information, especially if the data can be exported to an external graphical package.

Database Design

The first step in designing a database is to evolve from a single table containing all of the data items to be recorded to multiple tables with fewer columns and defined inter-relationships. A single table form of a database would contain unnecessary repetition of data with the possible danger of repeated data items containing inconsistent or anomalous values.

The database must be designed with no loss of data and in a form that allows tables to be merged to create different views of the data for users with different application or security access requirements.

Changes are easier to make in the development phase, so an iterative approach to database design involving the users is far more effective than waiting until the database is in full use.

A well-designed database can grow in a structured manner with the minimum of disruption to its users.

Normalisation

Normalisation techniques were originally designed and advocated by E.F. Codd in 1970 to decompose complex data structures into flat two-dimensional tables forming the basis of relational databases.

Normalisation provides anomaly-free data structures containing a minimum of redundant data. Because core items are only located in one table, amendments and updates are global, efficient, avoid inconsistency and reduce data storage requirements.

Basic Rules of normalisation :

- Choose unique primary key or combination of keys for every table - no duplicate rows, every row unique
- Remove repeating groups by reference to additional tables using foreign keys with codes corresponding to the repeating items allowing for global changes
- Ensure all non-key table columns are fully dependent on the primary key
- Introduce intermediate one:many tables to manage any many:many relationships encountered
- In general, normalisation should be balanced with database performance to ensure that information retrieval is optimised for the particular application required.

Data Quality

Data quality is an important consideration especially if the database information will be the basis of research publications.

Special consideration should be given to data at the point of entry to avoid the “garbage in, garbage out” syndrome :

- Use English UK version of software, ensuring that date formats are DDMMCCYY rather than the American version MMDDCCYY

Institute of Applied Health Sciences

University of Aberdeen

- Use range validation or “drop down” boxes on dates and numerical values to reduce data entry errors
- Use Referential integrity for reference tables whereby coded data cannot be entered in associated tables without a corresponding entry in the reference table and, similarly, a data item cannot be deleted from a table if it is referenced by another table within the database
- Minimise the use of free-text fields by using look-up tables where possible to decrease the possibility of inconsistency for extracting specific categories. For example, Positive or Negative could be recorded as (+, -), (+ve, -ve), (pos, neg), (POS, NEG), (Pos, Neg), (POSITIVE, NEGATIVE), (Positive, Negative) or (positive, negative) in free text, and this would make information retrieval more complex or even lead to incorrect analysis
- Meaningful names should be used for database tables and items to minimise errors when selecting items to formulate database queries
- Divide date information into separate fields comprising Day, Month and Century Year to provide faster access, where Year is likely to be used as a selector in database queries

High quality data will have the following characteristics :

- Information stored is complete, accurate and suitable to the current and potential future needs of the researcher
- Information can be easily retrieved in an efficient, timely, user-friendly manner
- Information is kept up to date with underlying changes – e.g. clinical classification changes, reorganisations, postcode updates, GP changes
- A high level of normalisation. reducing the possibility of inconsistent values appearing in fields within different tables.
- Referential integrity rules applied within the data management tools utilised preventing invalid input or deletion of referenced tables.
- Fully documented, reducing resource required for support

Performance

Although normalisation optimises data storage requirements, for a large database, this may cause degradation in retrieval times due to the additional work required rejoining multiple tables for particular queries. De-normalisation of some data items should be considered where particular tables are joined regularly and performance would benefit by re-introducing an element of redundancy. This would not normally be necessary for small databases.

Similarly, the introduction of indexes may speed up retrieval rates but at the expense of degradation of write rates on data entry, since there is an additional load to update indexes when new data is added to the database. Indexes should never be used where there is only a small number of rows.

Institute of Applied Health Sciences

University of Aberdeen

Database tuning is an on-going process and the basic design will need to evolve as the database is populated and the usage pattern becomes clear. Information retrieval performance may be enhanced if compound data items are split into separate data fields to allow more efficient data searches. For example, retrieval of patients living in a particular town, where data is held as an address string but would be more efficient split into separate address line and town fields.

Documentation

This aspect of database development is often neglected, but a well-documented system is essential to ensure continuity of database maintenance and development as experienced staff migrate to other projects.

Ideally, a User Manual should be produced to provide a high level description of the database function and usage in straightforward language and a Systems Maintenance Manual provided giving detailed information on the database structure, function, design, operational use and limitations.

All documentation should be dated, kept up-to date and preferably be held electronically on a centrally shared server.

Database Testing

Testing is a vitally important part of developing databases. Ideally, a small pilot database should be developed with a representative sample of the data sets and a test plan devised to cover all aspects of the database functionality. Potential users should be involved in the final testing phase to ensure that the design is suitable for their purpose. Only when the prototyping stage has been approved by the users and successful testing achieved, can the full database implementation be undertaken.

For Access databases, it is best to create and maintain tables in a development database, separate from the master database. Use of the “Linked Table Manager” facility will facilitate testing without any risk of corrupting the master data. The networked version of the program can be safely linked to the “actual” tables in another database.