Human Computer Interaction Development and Management

Tonya Barrier, Ph.D.
Southwest Missouri State University, USA
Chapter 16

Facilitating End User Database Development by Working with Users’ Natural Representations of Data

Valerie J. Hobbs and Diarmuid J. Pigott
Murdoch University, Western Australia

One of the main advantages of user-developed applications is considered to be the greater familiarity the users themselves have with the problem domain, and hence the greater likelihood of their creating an application that meets their needs. However, it is equally frequently reported that many end users lack the skills to develop applications that are of a high quality. Database modelling and relational database design, in particular, are known to be problematic for novices. We present two case studies in which the first stage of the development process was completed entirely by the end user, making use of their own understanding of the dataset, the problem domain, and tools that were familiar to them. In each case, they had represented the data in the form of lists. An IT expert then facilitated the conversion of the dataset to a relational database, with the participation of the end users throughout the process. The end users were able to see the concepts of database design emerge naturally from a problem that was already familiar to them, and to understand their importance in a practical manner.

Previously Published in Managing Information Technology in a Global Economy, edited by Mehdi Khosrow-Pour, Copyright © 2001, Idea Group Publishing.
INTRODUCTION

There is a large body of literature on end user computing and on user-developed applications (UDAs). UDAs are applications developed by end users, rather than IT professionals, to undertake some particular task. The basic premise of end user development is that the users themselves are in the best position to understand the requirements of the application domain and therefore to create an application tailored to their particular needs. Thus, one of the often-quoted advantages of end user application development is that it eliminates the problematic step of user requirements elicitation in the systems development lifecycle (Agboola, 1998). Indeed, Lally (1995) noted that a key issue in end user computing is determining how organizations can more effectively utilize the end user’s superior knowledge of the application requirements in the process of application development.

However, there is also a large body of literature pointing out the potential hazards of UDAs that are developed when users without adequate background or training develop their own applications. Lack of familiarity with development methodologies, or with application software, may result in significant errors in the final product, despite their understanding of the requirements (Panko & Halverson, 1996). Agboola (1998) attempted to separate the effects of modelling knowledge and application domain knowledge in an experimental setting, and found support for the commonly held view that application domain knowledge was less important than modelling knowledge as a predictor of database implementation correctness.

Database management systems, today overwhelmingly personal relational systems such as Microsoft AccessTM, FoxProTM or FileMaker ProTM, are one of the most common types of software used to create UDAs. However, normalised relational tables are not a natural way of representing data for most people. Further, relational model concepts such as keys, functional dependency and referential integrity, essential to create a correct and flexible database, are difficult to grasp. The spreadsheet format, where information is set out in tabular format, is generally thought to be more intuitive for novices, although it is certainly not immune from errors (Panko & Halverson, 1996).

Hutchins, Hollan and Norman (1985) postulated the concept of “gulf” between a user’s goals and the computer interface: if the gulf between the user’s goals and the computer interface is large, the user will require more effort to accomplish their goals. This gulf is made up of semantic distance, which is the gulf between a user’s conceptualisation of the “real world” object and an abstraction of it; and articulatory distance, which is the gulf between the meaning or abstraction of the object and its physical form or syntax. Batra (1993) further identified mapping (transforming the world into a representation), rules (which govern the mapping) and consistency (between the world and the representation) as distinct-
tions within semantic distance. Conceptual and logical database design, which involve translating the semantics of a problem into a data representation, are recognised to be difficult tasks for novice designers (Batra & Antony, 1994). This may be attributed to a large semantic distance between the user's view of the world and its representation as a data model or set of relational tables. Although current DBMS products such as Microsoft Access™ typically offer the end user a great deal of assistance in implementation of a design, relatively little help is offered in the earlier, modelling stage.

Work with novice and expert database designers has provided some insight into the types of errors that inexperienced data modellers typically make. Whereas novices can generally identify entities and attributes correctly, they frequently have problems with connectivity of relationships (e.g., Batra, Hoffer, & Bostrom, 1990), especially binary and ternary relationships. Batra (1993) reported that novices also had more problems representing binary relationships, particularly many-to-many, in relational tables than in Entity-Relationship diagrams.

Many authors (e.g., Kreie, Cronan, Pendley, & Renwick, 2000; Salchenberger, 1993) see training end users in analysis and design techniques as essential to improving UDA quality. Ahrens and Sankar (1993) recommended software tutors that teach specific database design skills. Batra and Antony (1994) suggested that an approach based on heuristics would be more efficient than one based solely on database principles. Batra (1993) developed a typology of database design errors, and suggested ways in which training strategies and knowledge-based tools could be developed to address particular types of errors. Agboola (1998) concluded that end users needed a large amount of hands-on database training to become competent, and suggested that collaborative development between end users and professionals could be a viable solution.

In this chapter, we describe two case studies of user-developed database applications that were facilitated by an IT professional. In each example, the users were experts in the application domain, and the data requirements were fully known. Both sets of users were computer literate, one with extensive spreadsheet experience, but neither had any database development experience and had called in one of the authors (DP) as a consultant when they realised database functionality would be required. The consultant made the decision at the outset to work directly with the users' own models of the data, and to convert them directly to database implementations with the active involvement of the users. One data set was in a series of spreadsheets, while the other was in the form of a table in a desktop publishing document. Throughout the process of conversion to the database the end users were instructed in the principles of database design, with relational model concepts emerging naturally from the need to keep the data consistent within its application domain. The aim of the consultant was to make use of the users' deep understanding of their particular dataset to facilitate their understanding of the
general database concepts. In this way it was hoped they would be left with the confidence and competence to be “self sufficient” in maintaining the database applications in future.

THE CASE STUDIES

Irish parish records

The first case study concerned historical research into the parish records for Killucan and Kinnegad in Co. Westmeath, Ireland. The research concerns “bringing back to life” a village that had lost most of its records, by assembling the sum of information available from various resources, including the baptismal registry and the occupancy of graveyards, Westmeath Grand Jury presentment books, and the vestry service records. A vast number of official records were destroyed during the Troubles in Ireland in 1922, making these isolated pockets of information highly important to genealogists.

The information had been collected in the field in notebooks by the historian, and had been transferred to embedded tables in Microsoft Publisher\textsuperscript{TM} documents. Publisher was originally chosen as the research was to be included in a book on reconstructing the identities of defunct villages.

The challenge of reconstruction was to identify individuals and their lives from the many different records and sources. The database was needed to assist the historian in finding possible matches and discrepancies in the various forms of data, before the data was imported to a genealogical program using the GEDCOM transfer format, and to improve the accuracy of the tables in the original Publisher document. The historian further realised that database functionality would be needed to be able to make certain queries, such as listing all the known births and deaths in a particular parish, or matching up deaths of mothers with births of children or remarriage of fathers. Tracing the occupations of an individual as recorded throughout his life could also provide some measure of social mobility.

Environmental assessment

The second case study was a database of ecological data, collected by an environmental consulting company (Ecoscape) as part of a series of vegetation surveys of reserves for the Western Australian Department of Conservation and Land Management (CALM). Environmental assessment studies were to be carried out for three shires (Mount Marshall, Lake Bryde and Kent), including identification of vegetation units; the distribution of vascular flora; identification of rare or threatened plant taxa; identification of species and communities of significance; analysis of the relationships between soil, landform and vegetation characteristics; and assessment of vegetation vigour and health, particularly in relation to rising water tables and salinity.
Information was to be collected both at the level of the administrative unit (the Crown land reserve) and at the level of the unit of ecological sampling (the quadrat) within those reserves. At the quadrat level the information included soil chemistry, soil landform data, crown cover data, and vegetation species data. At the reserve level, the information included more general information, such as adjacent and previous land use, water features, presence of industry, road access and amenities, sites of cultural significance, and reserve intactness.

The data requirements were strictly prescribed in the tender documents and it was essential for the finished database to integrate with the CALM Herbarium species lists and the Department of Land Administration GIS (Geographic Information Systems) datasets as well as other pre-existing State databases. The terms of the tenders required the delivery of the information as Microsoft Access 97™ databases.

The consultant was first called in at the start of the project. Since the users were very familiar with spreadsheets, but had no database experience at all, the consultant made the decision to do all the initial data input using spreadsheets. The worksheet format also had the advantage of more directly reflecting the data collection process (on field data sheets), and permitted a degree of flexibility in structuring the data. This was desirable since Ecoscape's experience had shown that various ambiguities and new requirements in field data were likely to necessitate the alteration of the worksheets on an ad hoc basis as the project progressed. It also enabled Ecoscape to settle on their own precision values for numbers (e.g., integer, float, etc.) and modify them on the fly, rather than having to define fixed data types in advance.

The Ecoscape project team designed and constructed the spreadsheets themselves in Microsoft Excel 97™, and filled them in using the data from the collection sheets. At this stage the consultant was called back to facilitate the conversion of these spreadsheets into a database format.

DEVELOPING THE DATABASES

The Process

The conversion of the lists to a relational database involved several stages. We summarise these stages next, before describing the details of the process for each of the case studies.

- **Step 1: Discussion.** The process commenced with an informal discussion between the consultant and the end users. The users described their data and the sorts of processing they needed to undertake. The consultant gave the users a brief outline of relational databases to explain how their lists would eventually become a set of tables that could be used for querying and reporting. This explanation included constructing a rough sketch of the proposed tables and
the relationships between them, to gain feedback from the users. The discussion stage was vital to ensure the users could clarify any issues with the data with the consultant and correct any misapprehensions he had, and for the users to be prepared for the remainder of the process.

- **Step 2: “Cleaning” the data.** This involved checking that all values were correct, and checking the interpretation of empty values in the lists (i.e., whether they represented “no value,” “value not known,” or a match of the value above). “No value” cells were labelled “n.v.” and “not known” were labelled “n.k.” This step was done in spreadsheet format for ease of data manipulation.

- **Step 3: Creating the equivalent of a First Normal Form relation (Date, 1991) in the spreadsheet.** This was done by using Fill Down where there was repeating data to ensure every cell had a value.

- **Step 4: Creating and populating the set of normalised relations (Date, 1991) in the database, directly from the spreadsheet.** This was done table by table, starting with “parent” tables, and involved:
  - Copying one table’s worth of data at a time to the database from the spreadsheet, to create a new table
  - Creating the primary key of the table
  - Creating (where necessary) the appropriate foreign key values from the data in previously created tables
  - Setting referential integrity between the new table and previously created tables
  - Altering the field length or data types where necessary

- **Step 5: Reconstitution.** Once all the tables were created and populated, queries were made to check that the normalisation had been lossless (Date, 1991) and to demonstrate to the user that all of the original data was still present in the normalised tables.

- **Step 6: Reviewing and formalising the database design process with the user.** The consultant used the Relationships window in Access to demonstrate to the user how the final tables linked together, and to discuss the concepts of keys and referential integrity more formally. The consultant then returned to the concept of data modelling using Entity-Relationship diagrams (Chen, 1976), introduced informally in Step 1, and demonstrated how a formal ERD would be drawn for the system just constructed. This led to discussion of how data modelling is used in the early stages of database design, and how the resulting relational design facilitates querying and reporting.

**IRISH PARISH RECORDS DATABASE**

The original data was in the form of tables embedded in a set of Microsoft Publisher documents, each document listing the records for a source of informa-
tion for a particular parish. The tables read left to right, with blank records indicating a repeat of the information listed above. Figure 1 shows part of the table for baptismal records from Killucan parish.

The process of converting to a database was then carried out following the steps described above.

- **Step 1: Discussion**

Several issues relating to the application domain were resolved during the initial discussion between the historian and the consultant. It was decided not to include the information for the vestry records and gravestones, since there was so little available, and to use only the baptism records. The consultant and historian discussed the possibility of recording multiple marriage in the database, but made the decision not to as the sources only recorded multiple marriages for men but not women. The absence of any remarriages in the small data set meant that a many-to-many relationship between the wives and husbands tables was avoided.

A preliminary diagram showing the proposed entities in the database was drawn and validated with the end-user. These entities were Fathers, Mothers, and Children.

- **Step 2: Data cleaning**

The original tables were copied from Publisher into Excel to facilitate data cleaning. All the baptism records from the two parishes were copied into a single worksheet, with an extra column “Parish” added to record the parish name implicit in the original document titles. As some wives had surnames as well as first names recorded a column “Wife other surname” was added to the worksheet,
Figure 2: 'First normal form' data for Killucan baptism records.

<table>
<thead>
<tr>
<th>Parish</th>
<th>Family Name</th>
<th>Birth Date</th>
<th>Occupation</th>
<th>Wife</th>
<th>Wife given name</th>
<th>Father</th>
<th>Father's given name</th>
<th>Sex</th>
<th>Parental Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Killucan</td>
<td>ASHE</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>AUSTIN</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>BRAZIER</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>BRAZIER</td>
<td>George</td>
<td>Butcher</td>
<td>Alice</td>
<td>William</td>
<td>27-Sep</td>
<td>1760</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>BRAZIER</td>
<td>George</td>
<td>Aescott</td>
<td>Elizabeth</td>
<td>Margaret</td>
<td>12-Aug</td>
<td>1745</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>BRAZIER</td>
<td>George</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>BRAZIER</td>
<td>George</td>
<td>n.k.</td>
<td>n.k.</td>
<td>n.k.</td>
<td>Richard</td>
<td>16-May</td>
<td>1700</td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>BRAZIER</td>
<td>George</td>
<td>n.k.</td>
<td>Catherine</td>
<td>Richard</td>
<td>18-Jun</td>
<td>1725</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>BURGES</td>
<td>n.k.</td>
<td>n.k.</td>
<td>Catherine</td>
<td>Alice</td>
<td>19-May</td>
<td>1702</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>CALDWELL</td>
<td>William</td>
<td>n.k.</td>
<td>Suzanne</td>
<td>RICHARDS</td>
<td>n.k.</td>
<td>n.k.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>CALDWELL</td>
<td>William</td>
<td>Blackwater</td>
<td>Elizabeth</td>
<td>Esther</td>
<td>06-Apr</td>
<td>1729</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>CALDWELL</td>
<td>William</td>
<td>Blackwater</td>
<td>Elizabeth</td>
<td>John</td>
<td>20-Feb</td>
<td>1730</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>CADD</td>
<td>George</td>
<td>Blackwater</td>
<td>Elizabeth</td>
<td>William</td>
<td>07-Feb</td>
<td>1740</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>CADD</td>
<td>George</td>
<td>Blackwater</td>
<td>Elizabeth</td>
<td>John</td>
<td>12-May</td>
<td>1743</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>CADD</td>
<td>George</td>
<td>Cusack</td>
<td>Elizabeth</td>
<td>John</td>
<td>16-Jun</td>
<td>1744</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>CADD</td>
<td>George</td>
<td>Harness Meker</td>
<td>Elizabeth</td>
<td>Esther</td>
<td>16-Sep</td>
<td>1745</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>CADD</td>
<td>George</td>
<td>James</td>
<td>Eleanor</td>
<td>Samuel</td>
<td>30-Aug</td>
<td>1759</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>CADD</td>
<td>George</td>
<td>James</td>
<td>Eleanor</td>
<td>Matthew</td>
<td>16-Sep</td>
<td>1710</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>CADD</td>
<td>George</td>
<td>James</td>
<td>Eleanor</td>
<td>Edward</td>
<td>30-Oct</td>
<td>1713</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Killucan</td>
<td>CADD</td>
<td>George</td>
<td>James</td>
<td>Eleanor</td>
<td>John</td>
<td>16-Jul</td>
<td>1716</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and the surnames moved to that column from the "Wife" column. Unknown or missing data was given "n.k." values.

- **Step 3: First Normal Form**

  All empty cells that indicated repeating data were filled in. Figure 2 shows the data set at the completion of this step.

- **Step 4: Normalised tables**

  The first table created was "Fathers." A new table was created in Access Datasheet view, and the contents of the worksheet columns "Parish," "Family name," "Father" and "Occupation" were pasted into this table using Paste Append. The fields were given the same name as the worksheet column names. From this temporary table the new table called "Fathers" was created from a make-table query retrieving all fields, using SELECT DISTINCT to ensure no repeating data. An AutoNumber field was added as primary key.

  A table called "Mothers" was created in a similar manner, based on the contents of the columns "Parish," "Family name," "Father," "Occupation," "Wife" and "Wife other name" in the worksheet, with an Autonumber primary key. The names of the columns in the worksheets were used for the field names. The field data types were all left as text.

  A field "HusbandFK" was added to the Mothers table to act as a foreign key to the Fathers table. The values for HusbandFK were added through an update query matching the Mothers and Fathers tables on the fields "Parish," "Family name," "Father" and "Occupation." The relationship between the two tables was created in the Relationships window, and referential integrity enforced. There were no errors. The fields "Parish," "Father" and "Occupation" were then deleted from the Mothers table, since they were only required for generating the foreign key
values. The table was deliberately kept in 2NF, as it was decided to retain the field “Family name” for ease of display.

Finally, a table called “Children” was created. This table was based on the spreadsheet columns “Parish,” “Family name,” “Father,” “Occupation,” “Wife,” “Children” and “Baptism dates.” As the baptism of a child is a unique event there was no need for the SELECT DISTINCT query here. Two new fields “FatherFK” and “MotherFK” were added to act as foreign keys to the Fathers and Mothers tables respectively. The fields FatherFK and “MotherFK” were populated using an update query as described above, and the redundant fields removed from the Children table. Again, the field “Family Name” was kept for the sake of convenience. Referential integrity was enforced between Children and Mothers, and Children and Fathers, with no errors. An AutoNumber field was added as a primary key, as there were 13 first and family name pairs repeated at least once in the dataset.

Figure 4: The Relationships window in Access showing final table design for the Parish records database.
• **Step 5: Reconstitution**

A query was then made to rejoin all records to show that there had been no loss of data from the original tabular data set (Figure 3). Filters were used to show the original two lists (Kinnegad and Killucan) as logically separate.

Several other queries were demonstrated to the user, based on her requirements. For example, the “Find Duplicates” query wizard was used to show repeated children’s name pairs, and to point out possible matches between people and occupations.

• **Step 6: Review**

The Access Relationships window was used to demonstrate how the final state of the database matched the rough diagram discussed with the historian at the start of the process (Figure 4).

**Environmental assessment database**

Prior to the consultancy, Ecoscape had designed the field data collection sheets in accordance with the tender documents. These data collection sheets were required to be kept as a permanent record. The data collection sheets were designed in sections, each section grouping together the different information required to be collected under the terms of the contract (Figure 5).

As the completed field sheets came back from the survey teams, together with the samples and site photographs, the data they contained was transferred to the spreadsheets designed by the Ecoscape project team. There were three separate Microsoft Excel 97 workbooks, one for each project (shire). Each workbook contained a number of worksheets for the different types of data collected. These worksheets corresponded to the sections on the field data collection sheets.
Figure 6: The 18 sheets of data collected for each shire.
- Reserve details and descriptions (Figure 7)
- Reserve locations
- Reserve water resources
- Reserve environmental weeds (Figure 8)
- Reserve rare flora
- Reserve commercial land use
- Reserve adjacent land use
- Reserve fauna
- Reserve amenities/tourism facilities
- Reserve sites of cultural significance
- Reserve PIN data
- Quadrat details and descriptions (Figure 9)
- Quadrat vegetation descriptions (Figure 10)
- Quadrat crown cover
- Quadrat crown cover gap and width measurements
- Quadrat soil and landform details
- Quadrat soil profile
- Quadrat soil chemistry

Figure 7: Reserves details worksheet for Mt. Marshall.

<table>
<thead>
<tr>
<th>Reserve No</th>
<th>Polygon IDNo</th>
<th>Reserve Name</th>
<th>Lot Number</th>
<th>Location</th>
<th>Reserve Type</th>
<th>Land District</th>
<th>Remarks</th>
</tr>
</thead>
</table>

Figure 8: Distribution of weeds in reserves for Mount Marshall.

<table>
<thead>
<tr>
<th>Reserve No</th>
<th>Reserve Name</th>
<th>Weed Species</th>
<th>Weed Infestation</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>12683</td>
<td>Warancing Nature Reserve</td>
<td>Ehena calycyna</td>
<td>4</td>
<td>Understory largely comprised of weeds</td>
</tr>
<tr>
<td>12689</td>
<td>Warancing Nature Reserve</td>
<td>Solarium holography</td>
<td>4</td>
<td>Understory largely comprised of weeds</td>
</tr>
<tr>
<td>VCL02</td>
<td>Vacant Crown Land 2</td>
<td>Echium plantagineum</td>
<td>2</td>
<td>Weed cover patchy</td>
</tr>
<tr>
<td>VCL02</td>
<td>Vacant Crown Land 2</td>
<td>Plantago sp1</td>
<td>2</td>
<td>Weed cover patchy</td>
</tr>
<tr>
<td>VCL02</td>
<td>Vacant Crown Land 2</td>
<td>Brassica touriitnrolli</td>
<td>2</td>
<td>Weed cover patchy</td>
</tr>
</tbody>
</table>

Figure 9: Quadrat details worksheet for Mount Marshall.

<table>
<thead>
<tr>
<th>Date</th>
<th>Collector</th>
<th>Reserve No</th>
<th>Region Code</th>
<th>Quadrat No</th>
<th>Datum</th>
<th>Zone</th>
<th>Easting</th>
<th>Northing</th>
<th>Accuracy</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.05.00</td>
<td>WEGJ</td>
<td>2523</td>
<td>738218</td>
<td>MM001</td>
<td>AG041</td>
<td>50</td>
<td>587881</td>
<td>6053493</td>
<td>3.3</td>
<td>50</td>
</tr>
<tr>
<td>13.05.00</td>
<td>WEGJ</td>
<td>2523</td>
<td>738218</td>
<td>MM002</td>
<td>AG041</td>
<td>50</td>
<td>5878127</td>
<td>6053870</td>
<td>4.7</td>
<td>45</td>
</tr>
<tr>
<td>13.05.00</td>
<td>WEGJ</td>
<td>2523</td>
<td>738218</td>
<td>MM003</td>
<td>AG041</td>
<td>50</td>
<td>5878396</td>
<td>6053924</td>
<td>3.8</td>
<td>60</td>
</tr>
<tr>
<td>13.05.00</td>
<td>WEGJ</td>
<td>2523</td>
<td>738218</td>
<td>MM004</td>
<td>AG041</td>
<td>50</td>
<td>5878156</td>
<td>6053924</td>
<td>5.8</td>
<td>45</td>
</tr>
<tr>
<td>13.05.00</td>
<td>WEGJ</td>
<td>2523</td>
<td>738218</td>
<td>MM005</td>
<td>AG041</td>
<td>50</td>
<td>5878280</td>
<td>6053424</td>
<td>3.7</td>
<td>45</td>
</tr>
<tr>
<td>13.05.00</td>
<td>WEGJ</td>
<td>2523</td>
<td>738218</td>
<td>MM006</td>
<td>AG041</td>
<td>50</td>
<td>5878280</td>
<td>6053424</td>
<td>3.7</td>
<td>45</td>
</tr>
<tr>
<td>13.05.00</td>
<td>WEGJ</td>
<td>2523</td>
<td>738218</td>
<td>MM007</td>
<td>AG041</td>
<td>50</td>
<td>5878280</td>
<td>6053424</td>
<td>3.7</td>
<td>45</td>
</tr>
<tr>
<td>13.05.00</td>
<td>WEGJ</td>
<td>2523</td>
<td>738218</td>
<td>MM008</td>
<td>AG041</td>
<td>50</td>
<td>5878280</td>
<td>6053424</td>
<td>3.7</td>
<td>45</td>
</tr>
</tbody>
</table>

Figure 10: Plant species for quadrats in Mount Marshall.

<table>
<thead>
<tr>
<th>Quadrat No</th>
<th>Species Name</th>
<th>Specie's Code</th>
<th>Voucher Number</th>
<th>Strata m</th>
<th>Growth Form</th>
<th>Average Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>MM001</td>
<td>Eucalyptus camaldulensis</td>
<td>EUCAL</td>
<td>MM00074</td>
<td>1</td>
<td>M</td>
<td>12.6</td>
</tr>
<tr>
<td>MM002</td>
<td>Eucalyptus camaldulensis</td>
<td>EUCAL</td>
<td>MM00074</td>
<td>1</td>
<td>M</td>
<td>12.6</td>
</tr>
<tr>
<td>MM003</td>
<td>Acacia clavigera</td>
<td>AGC1</td>
<td>MM00070</td>
<td>2</td>
<td>S</td>
<td>2.1</td>
</tr>
<tr>
<td>MM004</td>
<td>Eucalyptus camaldulensis</td>
<td>EUCAL</td>
<td>MM00074</td>
<td>1</td>
<td>M</td>
<td>12.6</td>
</tr>
<tr>
<td>MM005</td>
<td>Grevillea robusta</td>
<td>GREV</td>
<td>MM00075</td>
<td>3</td>
<td>S</td>
<td>5.6</td>
</tr>
<tr>
<td>MM006</td>
<td>Eucalyptus camaldulensis</td>
<td>EUCAL</td>
<td>MM00074</td>
<td>1</td>
<td>M</td>
<td>12.6</td>
</tr>
<tr>
<td>MM007</td>
<td>Eucalyptus camaldulensis</td>
<td>EUCAL</td>
<td>MM00074</td>
<td>1</td>
<td>M</td>
<td>12.6</td>
</tr>
<tr>
<td>MM008</td>
<td>Acacia clavigera</td>
<td>AGC1</td>
<td>MM00070</td>
<td>2</td>
<td>S</td>
<td>2.1</td>
</tr>
<tr>
<td>MM009</td>
<td>Eucalyptus camaldulensis</td>
<td>EUCAL</td>
<td>MM00074</td>
<td>1</td>
<td>M</td>
<td>12.6</td>
</tr>
<tr>
<td>MM010</td>
<td>Acacia clavigera</td>
<td>AGC1</td>
<td>MM00070</td>
<td>2</td>
<td>S</td>
<td>2.1</td>
</tr>
<tr>
<td>MM011</td>
<td>Eucalyptus camaldulensis</td>
<td>EUCAL</td>
<td>MM00074</td>
<td>1</td>
<td>M</td>
<td>12.6</td>
</tr>
<tr>
<td>MM012</td>
<td>Eucalyptus camaldulensis</td>
<td>EUCAL</td>
<td>MM00074</td>
<td>1</td>
<td>M</td>
<td>12.6</td>
</tr>
</tbody>
</table>

The worksheets are listed in Figure 6 and examples of data shown in Figures 7-10. Several of the worksheets recording reserve data had more than one record for each reserve (e.g., there were many fauna records for each reserve). Similarly, several quadrats had multiple records (e.g., there was more than one vegetation species per quadrat (Figure 10)).
The final dataset contained 2355 data points about 152 quadrats in 56 reserves. There were 946 data points collected about the reserves themselves.

The consultant then followed the same process to develop the Access database from the spreadsheets, with the participation of the Ecoscape project team.

**Step 1: Discussion**

The initial discussion identified a likely set of tables and the relationships between these tables were sketched. Since the information to be collected for each shire was largely the same and the same worksheets were found within each workbook, the data contained within the three workbooks could be combined in a uniform fashion.

It was clear that the main entities in the planned database were the “Project” (drawn from each workbook), “Reserve” (drawn from the Reserve Details worksheet in each work book) and “Quadrat” (drawn from the Quadrat Details worksheet in each workbook). The basic structure of the database was sketched showing each Project having many Reserves, and each Reserve having many Quadrats, and confirmed with the project team.

The names of the worksheets gave a clear indication of the levels at which the information was collected, and the sheets themselves were the obvious candidates for tables. Some were in a one-to-one but optional relationship with the reserves (e.g., commercial land use) or quadrats (e.g., photos), others were in a many to one (e.g., fauna for reserves, flora for quadrats). The difference between these types of relationships was discussed, together with the necessity for a parent record (i.e., a reserve or a quadrat) for each value held. (This led to some complications for interpolated data for some reserves at quadrat level in the data cleaning stage.)

The proposed tables containing the data from these worksheets were added to the diagram and confirmed with the team.

**Step 2: Data cleaning**

To record vegetation type data, a quadrat had to be created for each vegetation type for each reserve. Where this had not been done in the field (i.e., where the vegetation type had not been noticeably different and so had been missed) the information was added by interpolation using a GIS system. Vegetation type had already been added manually into the spreadsheet. This had necessitated the creation of “nominal quadrats” for vegetation data—uninvestigated areas of the Reserves that would be targets for inspection on the next field trip. Nominal quadrats records were therefore included in the “Quadrat Descriptions” worksheets.

The nature of blank cells was resolved (in consultation with Ecoscape) into those representing repeating data from the row above and those representing “no value.”
- **Step 3: First normal form**

  A “first normal form” worksheet was created from each of the original worksheets, using Fill Down function to ensure every cell had a value. An extra column was added to the “Reserve Details” worksheet in each of the workbooks, to hold the values for the Project IDs.

- **Step 4: Normalised tables**

  The first table to be created was “Projects,” to hold the information about the three shires. This data consisted of a code according to the CALM identification system, a reference title, and an automatically generated primary key. This table is parent to all the other tables.

  The data for Reserve Details and Descriptions was used to create a new table “Reserve Details.” The table was based directly on the spreadsheet data “Reserve Descriptions” as laid out, with the addition of a primary key for the table, and a foreign key to the project the reserve was part of. The fields were given the appropriate names according to the column headings in the workbook, and the table saved. Data for each of the three shires was added to the “Reserve Details” table in the same manner. Referential integrity was set with the “Projects” table.

  All the other worksheets for information pertaining to reserves were then converted into tables one at a time (i.e., the worksheet for “Water Resources” became a table called “WaterResources,” and so forth), following the same process. As each table was created and populated, referential integrity was enforced using the relationship manager. In the event that it was not possible to enforce referential integrity, the Find Unmatched Query Wizard was used to isolate the offending record. There were very few exceptions (about 5 for the entire project). These exceptions gave the consultant the opportunity to explain the importance of referential integrity in a very practical way.

---

*Figure 11: Union query for entire data set at reserve level.*
The "Quadrat" table was created next, using the same process as described for Reserves. Referential integrity for the quadrat-reserve relationship was ensured as the Reserve ID was already in place for each row in the Quadrat worksheets.

Finally, tables corresponding to the different information collected in Quadrats were created in the same manner as the various tables containing reserve information, and referential integrity set.

In some instances the data after normalising was multivalued: e.g., "traces of fauna" in a quadrat could record several traces - droppings, marks, carcasses etc. Based on the extent of its occurrence and likely significance in the system, the consultant and Ecoscape made the decision to represent the information simply as a list, rather than normalising further.

- **Step 5: Reconstitution**

The conversion process was finalised by creating queries to reconstitute each of the original 1NF worksheets. Based on these queries, a single union query was created to present all of the data at quadrat level in a single data set, and another
to provide all of the data at reserve level. Figure 11 shows part of the union query integrating all of the data at reserve level (for example, the last line visible in the table is derived from the worksheet of reserve weeds shown in Figure 8). This format was required by CALM in the terms of the tender, to interface with the GIS inquiry agent.

• **Step 6: Review**

The Access Relationships window was then used to show how the original proposed database diagram had been successfully implemented (Figure 12).

**DISCUSSION**

If the end users had followed a conventional process of database development (e.g., Connolly & Begg, 1998), they would most likely have analysed and modelled the data requirements (probably in an ERD), created the tables in the DBMS, and entered the data from the field collection sheets or notebooks (Figure 13a). As discussed earlier, this process is likely to involve a large semantic and syntactic “gulf” (Batra, 1993) between the user’s model of reality and the database implementation, with the possibility of errors and a poor quality product.

Instead, a two-stage process was followed. First, the end users developed their own systems for recording the required data in list format, using software with which they were familiar. In the second stage, the lists were converted into a
database by the consultant (Figure 13b).

The end users and the consultant were each able to work within their own area of expertise, reducing the semantic or syntactic distance they had to deal with. The end users, with their expert knowledge of the data, were able to work directly with the data and construct data sets that represented their view of the world in a very naturalistic manner. When the consultant entered the process, the data set was completely known – thus there were no problems of requirements elicitation apart from the clarification of a few minor points.

The facilitated process of converting from list format to database allowed relational database principles to be demonstrated to the users using data about which they were experts. We noted the natural emergence of abstractions from the data, which illustrated such concepts as entity, attribute, relationship, normalisation and referential integrity. For example, we found that the environmental assessment data had effectively been normalised by Ecoscape in the process of identifying the data collection points and recording in the worksheets, and the consultant had only to formalise the principle.

The consultant was able to follow a straightforward, structured process of converting the lists into a relational database. A factor in this was undoubtedly the help provided by the integration of the software packages used. Copying and pasting records from Excel to Access permitted an almost seamless transition between the spreadsheet and the database view of data. The ability of Access to provide autonumber keys, and the graphical representation of referential integrity, also made the definition of primary keys and foreign keys very straightforward. The Relationships window was valuable in helping the end users to see the ‘total’ picture and appreciate how queries and reports could be constructed.

The list format appears to be a more natural way of representing data for people unfamiliar with dealing with abstractions of data and logical relationships between them. The approach described here should be applicable to any other problem domain in which the end user is able to represent their data in the form of a list of facts with repeating values. (It should be noted that we are discussing only application domains suitable for representing as databases: “lists” requiring true spreadsheet functionality such as calculations would not be suitable.) The use of the list format as an intermediate modelling stage has inherent advantages in some situations: for example it was particularly suited to the fluid nature of the scientific data collected by the Ecoscape team.

Some minor disadvantages to the process were noted. Ecoscape had made use of cell comments, and this information was lost when the spreadsheets were converted. Any information about sequence implicit in the spreadsheet format would also be lost, as would intermediate data where only the final version of a derived value was copied (although neither applied in the examples here). To
retain these sorts of information would be possible, but would involve a less straightforward process of conversion.

We have not yet been able to follow up the progress of either the Ecoscape team or the historian, but it would be interesting to discover whether they are able to apply their new database skills to novel situations, such as continuing maintenance of this project, design and implementation of other projects, or (in the case of Ecoscape) instructing new members of their team. The participation of the end user and the explanation of the conversion process may impart sufficient understanding and skills for them to attempt similar projects in the future. However, it may be that the facilitation offered by the expert is an essential part of the process.

We believe that this approach of collaboration between end user and professional, making the best use of their separate areas of expertise, has great potential both to ensure high quality of database development and to provide appropriate training in database design for end users. The end users in the case studies reported were left with a sense of ownership of the process of design, and confidence in their ability to work with the finished product.

ACKNOWLEDGMENTS

We are grateful to David Kaesehagan, Sandy Griffin and Alan Hunter of Ecoscape for permission to use the Reserve case study, and to Sandra Pigott for use of the Parish records project.

REFERENCES


Wesley Longman.


