Nominal record linkage: the development of computer strategies to achieve the family-based record linkage of nineteenth century demographic data

Thesis

How to cite:


For guidance on citations see FAQs.

© 1989 The Author

Version: Version of Record
NOMINAL RECORD LINKAGE:
THE DEVELOPMENT OF COMPUTER STRATEGIES
TO ACHIEVE THE FAMILY-BASED RECORD LINKAGE
OF NINETEENTH CENTURY DEMOGRAPHIC DATA

by

JOHN ANTHONY WELFORD, B.Sc.

A Thesis Submitted
for
the Ph.D. Degree

Volume 2
of 2

Applied Historical Studies
The Faculty of Social Sciences
The Open University

Author's number: M7025784
Date of submission: 26th June 1989
Date of award: 31st October 1989
In this and the following chapter I come to the heart of the record linkage process: I shall be considering the provision of facilities for implementing the actual record linkage operations. My primary aim will be to examine the nature of the processes which are needed to retrieve individual source records from the Source Database and connect them in an appropriate fashion with others in a suitably structured Population Database.

In the last chapter I examined the processes which enable source records, as submitted by the user, to be converted into a more suitable form and then loaded into the Source Database. Within the overall view of the system I saw the prime function of these processes as being a means of transforming the records from a user-oriented form into a system-oriented form. In a similar fashion one can now choose to view the process of record linkage as being essentially a further stage in the transformation of the records.

In purely physical terms the transformation can be conveniently organised by transferring records from the Source Database to the Population Database. But, at the same time, there will be a need for the radical structural transformation of the individual records. In the Source Database each nominal record exists as a relatively isolated entity; but in the process of transferring the individual
records to the Population Database it will be necessary for them to be disaggregated and then re-aggregated into new forms and new structures. While the fundamental components of the Source Database are the individual nominal records, the fundamental components of the Population Database will be people, families and life-events.

In order to examine the process of family-based record linkage in terms of such record transformations I shall need to explore the following two major aspects of the problem:

1. the nature of the transformations, and, more specifically, how information should be organised in the Population Database.

2. the nature of the transformation processes, i.e. how records should be retrieved from the Source Database and restructured and linked in the Population Database.

In view of the complexity of the transformation processes which are involved in record linkage it will be appropriate to begin by presenting a preliminary analysis of the more significant strategic and design problems which must be dealt with. In the present system design the program which has responsibility for all linkage operations is the Record Linkage Subsystem, and in Section 6.5 I investigated its functional requirements and the characteristics of a suitable user interface. It was proposed that the process of linking nominal records should be accomplished by selectively transferring records from the Source Database and restructuring them in the Population
Database. And it was emphasized that in view of the current, exploratory nature of family-based linkage work the user should be given considerable control over the linkage operation and should be provided with extensive display facilities.

It is worth reflecting at this point that even before individual nominal records are submitted to the Record Linkage Subsystem some preliminary kinds of linkage will have already been accomplished. In the first place the translation of Christian names and surnames into their internal coded equivalents will have removed the problem of discrepant spellings: as a result two variants of the same name will be readily associated. And in the second place all the records in the Source Database are arranged into a network of clusters in such a way that it is possible to gain ready access to all the records containing one or more occurrences of a particular surname.

But despite this extensive 'preprocessing' and linking of the records the tasks which remain are significant ones. First, there is the need to supervise the organisation of records in the Population Database. Even with the moderately primitive structure which was illustrated in Figure 4.11 the creation and manipulation of family structures would present considerable organisational problems. When we come to consider a more realistic database structure we will find that it needs to be approximately twice as complicated as this, when measured in terms of the number of component record and set types, and so will be even more difficult to organise.
The second major task is the restructuring which needs to be applied to records when they are installed in the Population Database. Consider, for example, a typical census household record. In the Source Database such a record will have a very simple explicit structure, i.e. it will consist of one 'S-CENS-HOUSEHOLD' record and a number of associated 'S-CENS-PERSON' records. But such a household may also possess a very complex implicit structure, when one takes into account the kin relationships which can exist between the member occupants. And what needs to happen when the household record is transferred to the Population Database is that the implicit structure must be made explicit: this will require the location and/or creation of the necessary 'PERSON' and 'FAMILY' records and their associated sets so as to enable the required kin relationships to be modelled. In view of the wide variety and potential complexity of kin relationships this transposing of each household structure can clearly be a complicated operation.

The final significant linkage task is the one which is intrinsically associated with nominal record linkage, viz. the problem of matching records and deciding when linkage should take place. Although, as was suggested in Chapter 3, nominal ambiguity may be a reduced problem when linking nineteenth century records, this does not remove the requirement for making judgements about linkage, and so comprehensive matching strategies must still be provided.

Since the processes of record linkage represent the most complex part of the total system I shall, for simplicity, subdivide my
examination of them into three distinct parts. I shall begin by looking only at the record selection and display process, i.e. how particular groups of nominal records may be retrieved from the Source Database and printed. I shall then look in detail at the problem of structuring the linked data in the Population Database. Finally, in the next chapter, I shall proceed to consider the nature of the strategies which will enable groups of records to be linked and stored in the Population Database.

9.1 RECORD SELECTION AND DISPLAY

In Section 6.5 I argued that the user should be able to select a particular subgroup of records from the Source Database, say those for a particular surname, and initiate a linkage operation just on these. And I also argued that he should be able to have the records displayed so that he could then conveniently carry out his own manual linkage as a check against the system. What I propose to do in this section is to analyse the problems of selecting and displaying records and examine a suitable strategy for carrying out the necessary operations.

Figure 6.2 provided an example of record selection and display. In response to the command:

```
PP/PRINTSOURCES, LINK/BONE
```
the system searched the Source Database for records containing the surname 'BONE' or a synonym of it. As a result of this search it located and displayed five such records: three census records, one marriage record and a birth record.

The facilities described in Section 6.5 are designed to give the user considerable control over the selection process. At one extreme he can request to see particular types of records, e.g. death records, for selected surnames; at the other extreme he can request to see all the records in the database, printed in alphabetical order by surname.

In order to examine the processes which are involved in record selection and display I shall consider in detail a specific retrieval task, initiated by the user as follows:

PP/SOURCES/C5
PP/PRINTSOURCES/M

The first command specifies that the selection is to be restricted to 1851 census records (i.e. 'C5'), and the second command specifies that all those 1851 census records which contain one or more people whose surnames begin with the letter 'M' are to be displayed.

In the next section I shall consider how the required records in the Source Database are to be located and displayed. And in the following section I shall then look at the design of the part of the Record Linkage Subsystem which has responsibility for these processes.
9.1.1 Locating Records in the Source Database

In Section 8.2 I examined the processes which are involved in loading records into the Source Database and providing access to them via their component surnames. It is now necessary to consider the converse operation: the corresponding retrieval of records from the Source Database by the use of the surname access facility.

In order to accomplish the record selections specified in the retrieval example defined above two quite separate tasks must be performed. These are as follows:

1. For the specified surname initial character, viz. 'M', it is necessary to locate in turn all the surnames which begin with that character.

2. For each surname thus located it is necessary to use the index into the Source Database to locate the relevant census records.

Let us consider the first of these tasks. Information about surnames is kept in the Surname Directory, and Figure 9.1 shows the part of the Surname Directory which is needed to accomplish the first task. The first operation is to locate the 'N-SN-INITIAL' record corresponding with the letter 'M' by means of the 'CALC' mechanism. The members of the associated 'N-SN-ALPHA-SET' can then be retrieved in turn. Each 'N-SURNAME' member record will correspond with a
Implementation of the Record Linkage Subsystem: Ch. 9
part 1 - preliminaries

Figure 9.1 The Surname Directory: access to surnames via their initial letter (* = CALC record)

Figure 9.2 The Source Database: access to census records via surname (* = CALC record)
surname which begins with the letter 'M', e.g. 'METCALFE'. And from this record one can obtain the associated internal code, e.g. for 'METCALFE' this might be '10010401'.

In order to accomplish the second task it is necessary to use each surname internal code value as an index into the Source Database. The relevant records and sets involved in this operation are shown in Figure 9.2. Thus, for example, to locate all the 1851 census records for 'METCALFE' one must first use the 'CALC' mechanism on the code '10010401' to locate the corresponding 'N-SURNUM' record for 'METCALFE'. Counter fields within this record will indicate whether there are any census records connected, either within the 'S-CENS-NAME-SET' or within the 'S-CENS-LINK-SET'. Households in the former set will be those for which the head has the surname 'METCALFE'; households in the latter set will be those for which one or more occupants, but excluding the head, have the surname 'METCALFE'.

Assuming that there are member records in the 'S-CENS-NAME-SET' one then retrieves each 'S-CENS-HOUSEHOLD' record, and observes its type. Where this is not 'C5' (i.e. not 1851 census) it can be ignored; Where it is 'C5' the details are displayed, and the 'S-CENS-PERSON' records which are members of the associated 'S-CENS-OCC-SET' are retrieved in turn and displayed.

Finally, assuming that there are member records in the 'S-CENS-LINK-SET' one then retrieves each 'S-CENS-PERSON' record, and
observes its type. Where this is not 'C5' it can be ignored; where it is 'C5' it is then necessary to retrieve and display the whole household. This is achieved by first obtaining the owning 'S-CENS-HOUSEHOLD' record in the 'S-CENS-OCC-SET'. One can then display this record and the associated 'S-CENS-PERSON' member records, as in the previous case.

In this fairly restricted example the requirement was to display only those 1851 census records which contained one or more people whose surnames began with the letter 'M'. If, instead, it was required to produce an alphabetic listing for all 1851 census records then it would merely be necessary to carry out the operations described above for each letter of the alphabet in turn, starting at 'A' and finishing at 'Z'.

I have now established the nature of the mechanisms which enable selected records to be retrieved from the Source Database. I shall next consider how such records may be displayed. Since the records which are retrieved are in internal coded form, each code value needs to be translated back into its corresponding, original character string form prior to display: this operation will precisely reverse the actions carried out earlier by the Source Translation Subsystem.

Let us examine the process of converting name information held in coded form. This conversion is carried out by making use of information held in the Directory Database, and the relevant records and sets which are needed are illustrated in Figure 9.3. I have
Figure 9.3 The Directory Database: access to names via their internal code values (* = CALG record)
already examined this access route into the name directories in
Section 7.2: let us now merely consider a particular example of a code
conversion. Let us suppose, for example, that the code for
'ELWICK TOWNSHIP' is '00007002', and that it is required to convert
the code back into the name. Firstly, the 'CALC' mechanism is used in
association with the major code value '00007000' to obtain the
appropriate 'L-PN-NUMBER' record. This will be the owning record for
the synonym group in which 'ELWICK TOWNSHIP' is a member. All that
remains is to traverse the members of the 'L-PN-GROUP-SET' until the
record with code '00007002' is located: from this the corresponding
string 'ELWICK TOWNSHIP' can be obtained. The other four types of
names can be converted in a similar fashion.

Provision must also be made for the non-name fields in each
record to be converted into external form. Thus, for example, an
appropriate algorithm must be supplied to convert a computed date into
its equivalent calendar date form.

9.1.2 The Control Structure for Source Record Selection and Display

It is now necessary to take a preliminary look at the internal
design and control structure of the Record Linkage Subsystem.
However, for the present I shall be concerned only with those parts
which are responsible for overall control and for record selection and
display: the relevant modules are shown in Figure 9.4. The remaining
modules, which are responsible for record linkage operations, will be
Figure 9.4 The Record Linkage Subsystem - record selection and display: structure and flow of control
considered in the next chapter. Once again, as in earlier, similar diagrams the interfaces to the Input-Output Subsystem are not included.

The overall arrangement of the modules shown in Figure 9.4 has much in common with the corresponding one for the Source Loading Subsystem, shown in Figure 8.8. This similarity may be attributed to the complementary nature of their functions. Thus, the Source Loading Subsystem loads information into the Source Database; the modules shown in Figure 9.4 retrieve information from the Source Database. A marked point of similarity is the preponderence of database-handling modules: of the 18 modules shown in Figure 9.4, 14 interface directly with IDMS. Once again this reflects the functional bias of the modules, there being relatively little interaction with the user, but considerable interaction with both the Source and Directory Databases. The only major point of difference between the two subsystems does, in fact, concern access to the Directory Database. The Source Loading Subsystem has no requirement to access the Directory Database. The Record Linkage Subsystem, with its need both for extensive access to the Surname Directory and also for more general internal code to name conversion facilities, must access the Directory Database.

Let us now examine the flow of control through the modules shown in Figure 9.4. Entry to the subsystem is via the top module in the diagram, viz. 'M4RECORDLINK'. The module directly below this, 'PIENTRYPP', has the usual initialising and terminating functions. Thus, it is responsible for initiating the opening of the input and
output files and the database, and correspondingly at the end of the job for the closing of input and output files and the database. It also initiates the printing of opening and closing messages to the user. Finally, it has responsibility for processing the user's commands and for forwarding appropriate action requests to module 'PZDBPROCESS'. The range of commands available to the user was described in Section 6.5.

Module 'PZDBPROCESS' is responsible for supervising the opening of the database, the initialisation of the Population Database and the record display and linkage operations. For the present I shall concentrate attention only on the record display functions.

As was discussed in the last section there are two quite separate processes involved in record display:

1. sequencing through a series of surnames, which at one extreme can consist of just one surname, and which at the other can consist of all the surnames in the Surname Directory.

2. for each surname in the sequence, retrieving and displaying the types of records from the Source Database which the user has requested.

The sequencing of surnames is controlled by module 'DUSNSTREAM', and this operates in the following way. If, for example, the user has requested records for people whose surnames begin with the letter 'M',
then module 'PZDBPROCESS' calls 'DUSNSTEM' (1), and requests it to prepare to deliver the 'stream' of surnames constituting the range which has been requested. 'PZDBPROCESS' then calls 'DUSNSTEM' repetitively, to obtain each set of surname details (i.e. the name string plus internal code), until 'DUSNSTEM' reports back that the stream has been exhausted.

As 'PZDBPROCESS' obtains each surname it invokes 'X4SURPRINT' to display the surname and underline it with asterisks, as illustrated in Figure 6.2. It then uses the internal code value to retrieve the corresponding 'N-SURNUM' record in the Source Database. If the user has requested, for example, that only 1851 census records are to be displayed and the appropriate counters in the 'N-SURNUM' record indicate that census records do, indeed, exist for this surname then 'SVLINKCENSUS' is called. 'SVLINKCENSUS' will then supervise the subsequent handling of each census record. It traverses the two sets 'S-CENS-NAME-SET' and 'S-CENS-LINK-SET' (see Figure 9.2), and when it finds a household which is to be displayed (i.e. any 1851 household, in the present example) it invokes module 'SUCEPRINT' to supervise the translation and display functions.

Module 'SUCEPRINT' arranges to display the household details, and it also arranges to traverse the 'S-CENS-OCC-SET' so that each 'S-CENS-PERSON' record can be retrieved and displayed. For each internal code value which is to be translated into a name string it invokes module 'DVDBSEARCH' (2) to access the name directories and carry out the conversion. In a similar fashion, where computed dates
are to be converted into calendar dates module 'T3GETDATE' is invoked. For example, if 'T3GETDATE' were to be presented with the computed date code '62731183' it would return the string '2 OCT 1821 +- 183D'.

Birth, marriage and death records are retrieved and displayed in a similar fashion. For example, where marriage records are to be displayed the module 'USLINKMAR' traverses the 'S-MAR-NAME-SET', 'S-MAR-LINK-SET' and 'S-MAR-BR-LNK-SET' to gain access to the required marriage records in the Source Database (see Figure 8.6). As each marriage record is located module 'UTMARPRINT' is invoked to supervise the translation and display functions.

Finally, the closing of the database is carried out by module 'PYDBCLOSE'. This is called both at the end of a normal run and also when an irrecoverable error occurs in some database operation.

I have now shown how at the user's request individual records can be selected from the Source Database and displayed. In the remainder of this chapter I shall be concerned with the way in which such records, after appropriate transformation, can be organised in a suitably structured Population Database.
9.2 THE POPULATION DATABASE STRUCTURE

In Section 4.3.3 I explored the characteristics of a population database, starting from two primary requirements, viz. that such a database should:

1. be capable of modelling 'real-life' genealogical relationships: father, mother, cousin, step-son, etc.

2. be able to provide access to subgroups of the people and families in the database which are of potential interest, e.g. those people born in a particular year, those families for which the marriage was registered in a particular year, etc.

By the end of Section 4.3.3 I had developed a primitive population database structure, which was illustrated in Figure 4.11. This consisted of eight record types and ten set types. It was recognised that this structure had a number of limitations. In particular, it was noted that the simple name index which it employed could not cope with name variations, and, in particular, with a woman's change of name at marriage. It should also be observed that some of the components of this primitive population database, such as the 'JOB-LINK' and 'PLACE-LINK' records, were included in order to support particular kinds of database access, rather than because they corresponded with 'real world' entities. In the development of a more elaborate and practical scheme one would therefore wish to resolve these and similar limitations, and in the process create a structure
which can more accurately model the fundamental entities and structures which exist in the 'real world'.

The scheme which has been developed to satisfy these more stringent requirements is illustrated in Figure 9.5. Since this structure employs 17 record types and 17 set types it is approximately twice as complicated as our original primitive structure.

At the centre of the Population Database are the 'P-PERSON' and 'P-FAMILY' record types, together with their associated sets, which are used to model 'real-life' genealogical relationships. Below the 'P-PERSON' record in the diagram are the two record types and the associated 'P-LIFE-EVENT-SET', which are used to organise a 'life-history' for each person in the database. The other significant structural components provide the four major access routes into the database. The three records above the 'P-PERSON' record provide the Name Index, while the four records above the 'P-FAMILY' record provide the Cohort Index. Two other indexes are provided by records in the lower half of the diagram. Thus, the Occupation Index consists of the 'P-OCC-ENTRY' and 'P-OCC-CENS-ENTRY' records, and their associated sets, while the Place Index involves the corresponding use of the 'P-CENS-ENTRY' and 'P-CENS-HOUSEHOLD' records, and their associated sets.

In the following sections I shall examine these major structural components of the Population Database, and consider the design principles which are fundamental to their construction.
A complete set diagram of the Population Database

(* = CACR record)
9.2.1 Population Structures

In order to model 'real-life' genealogical relationships in the way that was described in Chapter 4 it is necessary to provide two distinct record types, one describing people and the other describing families. These record types, which I have called 'P-PERSON' and 'P-FAMILY', are illustrated in Figure 9.5. In addition one must provide the three set types which are used to connect together the 'P-PERSON' and 'P-FAMILY' records so that together they can model the corresponding 'real world' structures. These I have called 'P-FATHER-SET', 'P-MOTHER-SET' and 'P-CHILD-SET'.

Whereas the individual nominal records stored in the Source Database effectively provide 'snapshots' of events in the lives of persons and families, the function of the 'P-PERSON' and 'P-FAMILY' records in the Population Database is to model as closely as possible the corresponding 'real world' entities. Each 'P-PERSON' record is designed to act as a repository for information about a particular individual, and as record linkage proceeds and more details are provided, it becomes possible to accumulate more details in the 'P-PERSON' record. At the same time the contents of each 'P-PERSON' record can be used during the linkage process to enable matching to be carried out as each new record from the Source Database is presented for processing. The 'P-FAMILY' record is designed to operate in a similar way as a repository for the accumulating information about a male-female union, either legitimate or illegitimate, and, more generally, to facilitate the processes of family-based linkage.
It is necessary now to explore the kinds of information which need to be stored in the 'P-PERSON' and 'P-FAMILY' records. I shall begin with the 'P-PERSON' record. The requirement for a unique person identifier has already been noted in Section 6.6. In addition, there will need to be stored the key information about the person, such as his name, date of birth, and so on. Other attributes which one might consider storing are occupation and residence. However, these particular attributes present a problem since they can take many values for a particular individual, i.e. a person can change his occupation and/or residence many times throughout his life. If one were to store such information in the 'P-PERSON' record then it would be necessary to set aside space in each record for the maximum number of attribute values (perhaps 20) which might occur. A more appropriate solution is to provide a separate 'event' record to hold one set of the occupation-residence attribute values, and then allow each 'P-PERSON' record to own any number of these records in an associated set. As will be seen later this solution also facilitates the implementation of the Occupation and Place Indexes.

The 'P-PERSON' record can also be used to hold 'system' information, i.e. information which is required only for the internal working of the Record Linkage Subsystem. For example, a battery of 'linkage status' indicators, specifying whether the person has yet been identified in the 1851 census, in a baptism record, etc., can aid the Record Linkage Subsystem in its matching operations. In the absence of such indicators the subsystem would need instead to scan the event records when making its matching decisions.
I shall now identify the main components of a 'P-PERSON' record. They are as follows:

- a unique person identifier, e.g. 'P00031'.

- a number of linkage status indicators, specifying, in particular, the types of records in which the person has been identified. They would, for example, show whether the person has been located in the 1851, 1861 and 1871 censuses, whether their birth or baptism record has been located, and similarly for their death or burial record. Other indicators can show whether the person is linked to his family of origin, i.e. via the 'P-CHILD-SET', and also whether, and in what manner, he is linked to the Cohort and Name Indexes: the precise operation of these indexes will be discussed in the next two sections.

- the person's name in internal coded form, together with initials and the appropriate gender code. For a female the surname code stored will be the one which corresponds with her maiden surname, where known (3).

- birth information, and specifically the person's year of birth, his computed date of birth and the internal code for his birthplace (4). An indicator specifying whether he was born in the base, peripheral or none-base zones is included, and there is also an illegitimacy indicator.
- death information, and specifically the computed date of death and the internal code for the place of death. There is also a status indicator, showing, for example, whether this was an infant death, a child death, the death of an unmarried person, the death of a widow, etc. (5)

- counters, showing how many 'P-FAMILY' records are owned by the person in the 'P-FATHER-SET' or 'P-MOTHER-SET', and how many 'event' records are connected in the 'P-LIFE-EVENT-SET' (6).

Let us now explore the kinds of information which need to be stored in the 'P-FAMILY' record. Each such record is designed to hold information about a particular male-female union. Where this union is a legitimate one the 'P-FAMILY' record will need to contain information about the marriage and the offspring from the marriage; where the union is an illegitimate one it will contain essentially only information about the offspring.

Strictly, information about the parents and children does not need to be maintained in the 'P-FAMILY' record: it can always be obtained by locating the appropriate 'P-PERSON' records in the associated 'P-FATHER-SET', 'P-MOTHER-SET' and 'P-CHILD-SET'. For example, to find the number of children in a family it is necessary merely to count the number of 'P-PERSON' records in the 'P-CHILD-SET'. However, for many purposes it can be more convenient if the various attributes of the family, such as the number of children, are maintained in the 'P-FAMILY' record (7). Other useful attributes
include name information and the significant dates in the family's history, such as the date of the marriage and the dates of birth of the first and last children. It would be natural to expect information about the marriage itself to be represented by a corresponding 'event' for each of the marriage partners. However, as will be shown later, it can prove useful to have some of this information duplicated and stored in the 'P-FAMILY' record.

There are some additional information fields which are required in the 'P-FAMILY' record. As for the 'P-PERSON' record each 'P-FAMILY' record needs its own unique identifier: this requirement was noted in Section 6.6. Again, like the 'P-PERSON' record, the 'P-FAMILY' record can be used to hold 'system' information, i.e. information which can aid the operation of the Record Linkage Subsystem. Typical 'linkage status' indicators would, for example, specify whether the family has yet been located in the 1851 census and whether the corresponding marriage record has been found.

I shall now identify the main components of a 'P-FAMILY' record. They are as follows:

- a unique family identifier, e.g. 'F00007'.

- a number of linkage status indicators. They would, for example, show whether the family has been located in the 1851, 1861 and 1871 censuses, and whether the corresponding marriage record has been located. Other indicators show whether the
family is linked to the 'P-PERSON' records for the husband and wife, i.e. via the 'P-FATHER-SET' and 'P-MOTHER-SET', respectively, and also whether, and in what manner, it is linked to the Cohort and Name Indexes. Again, the precise operation of these indexes will be discussed in the next two sections.

- name information for the husband and wife. For the husband there is his complete name in internal coded form, together with initials; for the wife there is only the internal code for her forename, together with initials. It was decided that the code for a woman's maiden surname should be stored only once, in her 'P-PERSON' record. The code for each of her married surnames will then be as is specified for each of her husbands in the corresponding 'P-FAMILY' records.

- marriage information, and, in particular, those items concerned with the bride and groom's age and marital status. Where there is evidence that the 'family' is an illegitimate union this will also be indicated.

- a family history, identifying in computed date form, where known, the dates of the beginning and end of the marriage, and the dates of birth of the first and last children. Such information is of demographic significance and is also of value when record linkage operations are in progress. Thus, if the end date of the marriage is known, e.g. through knowing
the date of death of the first partner to die, then the 
linkage of births beyond this date can be avoided.

- counters, showing how many sons and daughters there are in the 
  family.

- indicators, showing whether the husband and wife were born in 
  the base, peripheral or non-base zones.

I shall now briefly examine the three set types which are used to 
connect 'P-PERSON' and 'P-FAMILY' records, viz. 'P-FATHER-SET', 
'P-MOTHER-SET' and 'P-CHILD-SET'. Although these sets have been 
discussed in detail in Section 4.3.3 I shall at this stage expand a 
little on the way in which the set connections are manipulated.

The three set types provide the three basic kinds of genealogical 
connections: father-family, mother-family and child-family. And these 
three types of connections, when used in association, enable the 
'P-PERSON' and 'P-FAMILY' records in the Population Database to be 
linked into appropriate genealogical structures or family-trees. The 
use of the three sets in this way is therefore fundamental to the 
method of modelling population structures in a CODASYL-type database. 
The actual process of linking records together is carried out by the 
database system itself, although program logic within the Record 
Linkage Subsystem must decide which records are to be linked. For 
each required inter-record link the subsystem will issue a DML 
'CONNECT' instruction, and in response IDMS will make the appropriate
Finally in this section it will be appropriate to consider the method of allocating unique person and family identifiers. The scheme is implemented in precisely the same way as was adopted for providing identifiers for source records in the Source Database. The record type used in this case is 'P-ITEM-CNT', and within the Population Database there will be two occurrences of this record needed, one for the 'P-PERSON' identifiers and the other for the 'P-FAMILY' identifiers. When the database is created the two records are given initial values of '1', and as each 'P-PERSON' or 'P-FAMILY' record is created the appropriate 'count' value is used to construct a unique identifier and is then incremented by '1'. For example, the first 'P-PERSON' record to be created is given the value 'P00001', the second 'P00002', and so on.

The 'P-ITEM-CNT' record has an additional use. In Section 6.6 it was suggested that the user might wish to obtain brief statistical information about the contents of the Population Database, such as how many people and families are present and how many names are represented. In order to maintain such information five additional 'P-ITEM-CNT' records are used, and the items which are counted are: persons, families, surnames, person names and family names (9). When the Population Database is created the five records are given initial values of '0', and as each relevant item is created the corresponding count record is incremented by 1. An example of the way in which these counts are displayed was given at the end of Chapter 6.
9.2.2 The Cohort Index

In the previous section I examined a practical scheme for modelling 'real-life' genealogical relationships and structures in a population database. What I shall move onto in this and the following sections is the provision of appropriate facilities for accessing subgroups of people and families in the population database which are of potential interest (10). In this section I shall be concerned with the Cohort Index.

The primitive population database structure illustrated in Figure 4.11 had just one record type, viz. 'COHORT', to provide access to particular cohorts of people and families. This design enabled one to select a particular year, say '1771', obtain the corresponding 'COHORT' record, and then using the 'PERSON-COHORT' set locate all the people born in that year. And using the 'FAMILY-COHORT' set one could locate all the 'FAMILY' records for marriages registered in that year.

In Section 6.6 rather more elaborate selection criteria were proposed. Let us consider first the access to 'P-PERSON' records. In addition to being able to locate the 'P-PERSON' records for a particular year the more advanced scheme would enable records for a range of years to be located. At the upper limit of this range access would be provided to all the 'P-PERSON' records in the Population Database. The following selection criteria were also introduced:
1. selection by gender.

2. selection by birthplace zone code, where the code is either 'B', for the base zone, or 'N', for the non-base zone (11).

Provided with such criteria it should therefore be possible, for example, to request access to the 'P-PERSON' records for all male natives of the base zone who were born in the period 1811-20. A similar request to gain access to the comparable group of males who were born in the period 1821-30 could also be made, and the characteristics of the two groups could then be compared.

The provision of these more elaborate selection criteria can be conveniently accomplished by means of a two-level index. Thus, given a particular combination of gender and birthplace zone criteria one should be able to use the higher level index to gain access to all the people in the database sharing these criteria. Alternatively, using the lower level index one should be able to gain access to any subset of this group born in a particular year. Such indexing facilities can be provided by the use of two record types. These are as follows:

1. P-ALL-PERSONS. Within the Population Database only five occurrences of this record are needed, and each of these is able to provide access to a major subgroup within the population. The five subgroups are as follows:
The first four subgroups contain those people whose gender, date of birth and birthplace are known: such information will normally have been obtained from a baptism or census record. The fifth is a 'catch-all' subgroup, containing the remaining people: all those for whom there is incomplete information concerning gender, date of birth and birthplace (12).

After we have selected a particular major subgroup in which we are interested and have located the corresponding 'P-ALL-PERSONS' record access to the individual 'P-PERSON' records is achieved by moving to the lower level index. This may be accomplished by accessing the 'P-PERSON-COHORT' records in the 'P-PER-COHORT-SET'.

2. P-PERSON-COHORT. Let us consider the 'P-PERSON-COHORT' records which are owned by the 'P-ALL-PERSONS' record for male natives of the base zone. There will exist in the database
one 'P-PERSON-COHORT' record for those members of this major subgroup who were born in a particular year. For example, there will be a 'P-PERSON-COHORT' record for the male natives of the base zone who were born in 1831, and this record will own the corresponding 'P-PERSON' records in the 'P-PERSON-SET'. Therefore to locate every male native of the base zone it would be necessary to take in turn each of the 'P-PERSON-COHORT' records in the appropriate 'P-PER-COHORT-SET', and locate all its member 'P-PERSON' records in the 'P-PERSON-SET'.

When each new 'P-PERSON' record is created in the Population Database it will be necessary for it to be connected to the Cohort Index. This will require it to be inserted in the appropriate 'P-PERSON-SET', and, where necessary, the corresponding 'P-PERSON-COHORT' record will need to be created (13). A single 'P-PERSON-COHORT' record is set aside for all the persons in the fifth, 'catch-all' subgroup, since in the majority of cases the year of birth will not be known, and so a lower level of indexing will not be possible for them.

In order that the Cohort Index can function as a two-level index it is necessary for both the 'P-ALL-PERSONS' and 'P-PERSON-COHORT' records to be designated as CALC records. This is as indicated in Figure 9.5. Thus, for example, to locate the 'P-ALL-PERSONS' record for the male natives subgroup it is necessary to carry out a CALC
operation using together the codes 'M' and 'B', i.e. using the combined code 'MB'. While to locate the 'P-PERSON-COHORT' record for male natives born in 1831 it is necessary to carry out a CALC operation using together the codes 'M' and 'B' and the string '1831', i.e. using the combined code 'MB1831'.

Let us consider now how 'P-FAMILY' records should be made accessible from the Cohort Index. It was proposed that the access facilities should roughly parallel those provided for the 'P-PERSON' records. Thus, it should be possible to gain access to the 'P-FAMILY' records for a particular year, where in this case the year refers to the year of marriage registration. It should also be possible to obtain access to the 'P-FAMILY' records for a range of years, and at the upper limit this should extend to include all the 'P-FAMILY' records in the Population Database. To correspond with the birthplace selection criteria for 'P-PERSON' records it was proposed that similar selection criteria for 'P-FAMILY' records should be adopted, these to be based on the birthplace zone codes of both of the marriage partners. This would then facilitate the comparison of native and non-native families, a native family being defined in this context as one in which either or both of the marriage partners were born in the base zone.

The provision of these more elaborate selection criteria can be conveniently accomplished by means of a two-level index, similar to the one described above for 'P-PERSON' records. Thus, given a particular combination of birthplace zone codes one should be able to
use the higher level index to gain access to all the families in the database sharing these criteria. Alternatively, using the lower level index one should be able to gain access to any subset of this group for which the marriages were registered in a particular year. Such indexing facilities can be provided by the use of the following two record types:

1. P-ALL-FAMILIES. Within the Population Database only three occurrences of this record are needed, and each of these is able to provide access to a major subgroup within the population. The three subgroups are as follows:

   - families in which either or both of the marriage partners are natives of the base zone
   - families in which neither of the marriage partners are natives of the base zone
   - families which cannot be located in either of the above categories

The first two subgroups contain those families for which the marriage date and the birthplaces of both marriage partners are known: a typical example of such a family would be one that has been located in the 1851 census and whose marriage record is available. The third subgroup contains the remaining families, those for which the marriage date or the
birthplace information of the marriage partners is not available: a typical example of such a family would be one where the marriage partners have been identified only as parents in an isolated baptism record (14).

Access to the individual 'P-FAMILY' records via the Cohort Index operates in approximately the same way as access to the individual 'P-PERSON' records. We would therefore initially locate the 'P-ALL-FAMILIES' record which corresponded with the major subgroup in which we were interested. Access to the required 'P-FAMILY' records would then be achieved by moving to the lower level index. This would be accomplished by accessing the 'P-FAMILY-COHORT' records in the 'P-FAMILY-SET'.

2. P-FAMILY-COHORT. Let us consider the 'P-FAMILY-COHORT' records which are owned by the 'P-ALL-FAMILIES' record for families in which neither of the marriage partners are natives of the base zone. There will exist in the database one 'P-FAMILY-COHORT' record for those families in this major subgroup whose marriages were registered in a particular year. For example, there will be a 'P-FAMILY-COHORT' record for the non-base zone families whose marriages were registered in 1831, and this record will own the corresponding 'P-FAMILY' records in the 'P-FAMILY-SET'. Therefore to locate every family for which neither of the marriage partners are natives of the base zone it would be necessary to take in turn each of
the 'P-FAMILY-COHORT' records in the appropriate 'P-FAM-COHORT-SET', and locate all its member 'P-FAMILY' records in the 'P-FAMILY-SET'.

When each new 'P-FAMILY' record is created in the Population Database it will be necessary for it to be connected to the Cohort Index. This will require it to be inserted in the appropriate 'P-FAMILY-SET' (15). A single 'P-FAMILY-COHORT' record is set aside for all the families in the third, 'catch-all' subgroup, i.e. for those with missing marriage date and/or birthplace information.

As is indicated in Figure 9.5 both the 'P-ALL-FAMILIES' and 'P-FAMILY-COHORT' records are designated as CALC records: this provides the two-level indexing capability. Thus, for example, to locate the 'P-ALL-FAMILIES' record for the subgroup of families in which neither of the marriage partners are natives of the base zone it is necessary to carry out a CALC operation using the zone code 'N'. While to locate the 'P-FAMILY-COHORT' record for the members of this subgroup who were married in 1831 it is necessary to carry out a CALC operation using together the zone code 'N' and the string '1831', i.e. using the combined code 'N1831'.
The primitive population database structure illustrated in Figure 4.11 had just one record type, viz. 'NAME', to provide access to people with a particular name. This scheme had a number of fundamental limitations: for example, it was unable to cope with a woman's change of name at marriage. What must now be explored, therefore, is the design of a Name Index which is sufficiently elaborate to handle all of the requirements of the record linkage and population analysis processes.

The Cohort Index is provided almost exclusively for the benefit of the user: to enable him to gain access to particular groups of people and families which may be of interest to him. The Name Index, by contrast, is needed primarily by the Record Linkage Subsystem itself in order to carry out its linkage operations. For example, when presented with a burial record, say, for 'GEORGE WILSON' the Record Linkage Subsystem must access the Population Database in order to discover whether there are any 'P-PERSON' records already present for 'GEORGE WILSON'. Similarly, when presented with a baptism record, say, for a child of 'GEORGE & ISABEL WILSON' the Record Linkage Subsystem must access the Population Database in order to discover whether there are any 'P-FAMILY' records already present for 'GEORGE & ISABEL WILSON'. Such operations will require the availability of appropriate name indexing facilities.
The Name Index is also, however, of potential value to the user. In the first place he may use it to obtain access to particular individuals and families in which he is interested. And secondly he may use it when he wishes to assess the results of a particular linkage operation. Having, say, linked all the records for 'SMITH' he may wish to see displayed all the 'SMITH' people and families. The Name Index should be able to provide him with such access facilities as these.

I shall now proceed to identify the specific requirements which the Name Directory should be able to satisfy. They are considered to be as follows:

1. Given a particular male name, e.g. 'JOHN SMITH', there should be access to all the 'P-PERSON' records for males with that name.

2. Given a particular female name, e.g. 'ANN SMITH', there should be access to all the 'P-PERSON' records for females for whom this is the maiden name or the married name.

3. Given the names of marriage partners, e.g. 'JOHN & ANN SMITH' there should be access to all the 'P-FAMILY' records for marriage partners with these names.

4. At a more global level, given a particular surname, e.g. 'SMITH', there should be access to all the 'P-PERSON' records
for people with that surname (16) and access to all the 'P-FAMILY' records for families with that surname.

5. For all accesses the index should treat each major Christian name or surname as equivalent to its synonyms. Thus, for example, if the user submits the name 'ANN SMITH' and there is a 'P-PERSON' record present for 'HANNAH SMYTHE' then one would expect the index to locate this record.

The requirements for global searching (via surname) and more local searching (via individual name) would suggest that a two-level index is needed, as for the Cohort Index. Let us consider this in relation to the access to 'P-PERSON' records. The following two records can enable access at the two levels to be readily accomplished:

1. P-SURNAME. This record type corresponds closely with the 'N-SURNUM' record type in the Source Database, and it provides global access to 'P-PERSON' records. Within the Population Database there needs to be one 'P-SURNAME' CALC record for each major surname internal code value. Thus, for example, if the internal code value for 'SMITH' is '10015300' then the corresponding 'P-SURNAME' record would be located by using the 'CALC' mechanism on the major code value '100153'. For convenience and also efficiency each 'P-SURNAME' record contains the corresponding character string for the major surname (17). Once a particular 'P-SURNAME' record has been
located access to the individual 'P-PERSON' records is achieved by moving to the lower level index. This is accomplished by accessing the 'P-PERSON-NAME' records in the 'P-PER-NAME-SET'.

2. P-PERSON-NAME. Within the Population Database there will exist one 'P-PERSON-NAME' CALC record for each required full name. For example, there might be a 'P-PERSON-NAME' record for 'JOHN SMITH' and another one for 'ANN SMITH', and both of these would be connected in the same 'P-PER-NAME-SET', with the 'SMITH' 'P-SURNAME' record as owner. If the internal code values for 'JOHN' and 'SMITH' are '00003300' and '10015300', respectively, then the corresponding 'P-PERSON-NAME' record will be located by using the 'CALC' mechanism on the combined major codes '000033' and '100153', i.e. on the code '000033100153'. For convenience each 'P-PERSON-NAME' record contains the appropriate character string for the major Christian name.

Each 'P-PERSON-NAME' record can be the owner in two sets, 'P-BIRTH-NAME-SET' and 'P-WIFE-NAME-SET', although the second of these will always be empty for a male name. Consider first the 'P-PERSON-NAME' record for 'JOHN SMITH'. In the 'P-BIRTH-NAME-SET' it will own all the 'P-PERSON' records for people called 'JOHN SMITH'; it will have no members in the 'P-WIFE-NAME-SET'. Consider secondly the 'P-PERSON-NAME' record for 'ANN SMITH'. In the 'P-BIRTH-NAME-SET' it will own
all the 'P-PERSON' records for females whose maiden name is 'ANN SMITH'; in the 'P-WIFE-NAME-SET' it will own all the 'P-FAMILY' records for those families in which the wife's married name is 'ANN SMITH'. In order to locate such a wife's 'P-PERSON' record it is necessary merely to locate the owner of the 'P-FAMILY' record in the corresponding 'P-MOTHER-SET'.

In passing, it should be observed that the 'P-FAMILY' record is being used here as a 'link' record to support the n:n relationship which exists between women and names. Each woman can have several names; for each name there can be several women. Therefore an n:n relationship is required between 'P-PERSON-NAME' and 'P-PERSON'. This is achieved by using 'P-FAMILY' as the link record: it exists as a member of a set, 'P-WIFE-NAME-SET', in which 'P-PERSON-NAME' is the owner, and it simultaneously exists as a member of a set, 'P-MOTHER-SET', in which 'P-PERSON' is the owner (18).

Let us consider next the access to 'P-FAMILY' records. The provision of a two-level index can again be readily accomplished by the use of two types of records as follows:

1. P-SURNAME. This record has already been described above. It will be observed in Figure 9.5 that this record can simultaneously be the owner in two set types, 'P-PER-NAME-SET' and 'P-FAM-NAME-SET'. This position enables it to provide access to both 'P-PERSON' and 'P-FAMILY' records. In the
latter case after a particular 'P-SURNAME' record has been selected and obtained access to individual 'P-FAMILY' records is achieved by first accessing the 'P-FAMILY-NAME' records in the 'P-FAM-NAME-SET'.

2. P-FAMILY-NAME. Within the Population Database there will exist one 'P-FAMILY-NAME' CALC record for each required set of marriage partner names. For example, there might be a 'P-FAMILY-NAME' record for 'JOHN & ANN SMITH'. In this case if the internal codes for 'JOHN', 'ANN' and 'SMITH' are '00003300', '00000700' and '10015300', respectively, then the corresponding 'P-FAMILY-NAME' record will be located by using the 'CALC' mechanism on the combined major codes '000033', '000007' and '100153', i.e. on the code '000033000007100153'. For convenience each 'P-FAMILY-NAME' record contains the corresponding character strings for the major Christian names of the husband and wife.

Each 'P-FAMILY-NAME' record is the owner of a set, 'P-COUPLE-NAM-SET', and the members of this set are the 'P-FAMILY' records for all those pairs of marriage partners with the corresponding names.

A final observation which should be made about the Name Index concerns the creation of its constituent records. Since the index is required to contain the names of persons and families actually present in the Population Database, it follows that its constituent records
will need to be created dynamically, in parallel with the record linkage operations. When the Population Database is initially created, therefore, there will be no records in the Name Index. The same comments apply to the Occupation and Place Indexes, which are discussed later in the chapter.

9.2.4 Life Events

As a particular linkage operation progresses an increasing amount of information will be accumulated about the people in the Population Database. When, for example, a baptism record is linked then there will normally be a gain in information about three people: the father, the mother and the child. In the majority of cases we will learn at the bare minimum that the three people were alive on the date in question (19). We will also often have an indication of their place of residence, and for the father the occupation. For a legitimate baptism we will also have by implication the marital status of the father and mother.

This accumulating information can serve two purposes. In the first place it can assist the record linkage operation itself. If, for example, we know that on a particular date a person had the marital status 'married' then this can prevent us from linking to that person any candidate record for a later date in which the potentially linkable person has the marital status 'single'. In the second place the accumulating information can have an important substantive value,
in that it can serve to provide as complete a picture as possible of the individual life experiences of those present in the Population Database.

The question of how this information should be stored has already been discussed in Section 9.2.1. In view of the fact that it might be necessary to store a large amount of information about any one individual it was concluded that space should not be set aside for this in each 'P-PERSON' record: a more appropriate solution is to provide a separate 'event' record and allow each 'P-PERSON' record to own any number of these records in an associated set (20). Each event to be recorded will derive from some nominal record in which the person is mentioned, and the following types of events are recognised within the present system:

- census (1851, '61 and '71)
- birth or baptism
- birth or baptism (father and mother)
- marriage (21)
- death or burial

Thus, for example, in the Population Analysis listing which appears in Figure 9.6 there are provided the details of eleven events in the life
### Implementation of the Record Linkage Subsystem:

#### Part 1 - Preliminaries

**Figure 9.6** The organisation of a life-history in the Population Database: eleven events in the life of 'FRANCES WILSON'

<table>
<thead>
<tr>
<th>FATHER</th>
<th>Thomas Wilson</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOTHER</td>
<td>Ann Patterson</td>
</tr>
<tr>
<td>HUSBAND</td>
<td>Robert Pattison</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EVENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1822 BAPTISM:</td>
</tr>
<tr>
<td>30 SEP 1822</td>
</tr>
<tr>
<td>1849 MARRIAGE:</td>
</tr>
<tr>
<td>21 APR 1849</td>
</tr>
<tr>
<td>1853 MOTHER/BAP:</td>
</tr>
<tr>
<td>4 JAN 1853</td>
</tr>
<tr>
<td>1854 MOTHER/BAP:</td>
</tr>
<tr>
<td>11 APR 1854</td>
</tr>
<tr>
<td>1855 MOTHER/BAP:</td>
</tr>
<tr>
<td>12 AUG 1855</td>
</tr>
<tr>
<td>1857 MOTHER/BAP:</td>
</tr>
<tr>
<td>28 NOV 1857</td>
</tr>
<tr>
<td>1860 MOTHER/BAP:</td>
</tr>
<tr>
<td>11 APR 1860</td>
</tr>
<tr>
<td>1861 CENSUS:</td>
</tr>
<tr>
<td>7 APR 1861</td>
</tr>
<tr>
<td>1862 MOTHER/BAP:</td>
</tr>
<tr>
<td>24 AUG 1862</td>
</tr>
<tr>
<td>1864 MOTHER/BAP:</td>
</tr>
<tr>
<td>10 JUL 1864</td>
</tr>
<tr>
<td>1871 CENSUS:</td>
</tr>
<tr>
<td>2 APR 1871</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SPONS</th>
<th>7 OCT 1822</th>
<th>31 JUL 1778</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAUGHTERS</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FAMILY PO0038</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUSBAND</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EVENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1822 BAPTISM:</td>
</tr>
<tr>
<td>30 SEP 1822</td>
</tr>
<tr>
<td>1849 MARRIAGE:</td>
</tr>
<tr>
<td>21 APR 1849</td>
</tr>
<tr>
<td>1853 MOTHER/BAP:</td>
</tr>
<tr>
<td>4 JAN 1853</td>
</tr>
<tr>
<td>1854 MOTHER/BAP:</td>
</tr>
<tr>
<td>11 APR 1854</td>
</tr>
<tr>
<td>1855 MOTHER/BAP:</td>
</tr>
<tr>
<td>12 AUG 1855</td>
</tr>
<tr>
<td>1857 MOTHER/BAP:</td>
</tr>
<tr>
<td>28 NOV 1857</td>
</tr>
<tr>
<td>1860 MOTHER/BAP:</td>
</tr>
<tr>
<td>11 APR 1860</td>
</tr>
<tr>
<td>1861 CENSUS:</td>
</tr>
<tr>
<td>7 APR 1861</td>
</tr>
<tr>
<td>1862 MOTHER/BAP:</td>
</tr>
<tr>
<td>24 AUG 1862</td>
</tr>
<tr>
<td>1864 MOTHER/BAP:</td>
</tr>
<tr>
<td>10 JUL 1864</td>
</tr>
<tr>
<td>1871 CENSUS:</td>
</tr>
<tr>
<td>2 APR 1871</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EVENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1822 BAPTISM:</td>
</tr>
<tr>
<td>30 SEP 1822</td>
</tr>
<tr>
<td>1849 MARRIAGE:</td>
</tr>
<tr>
<td>21 APR 1849</td>
</tr>
<tr>
<td>1853 MOTHER/BAP:</td>
</tr>
<tr>
<td>4 JAN 1853</td>
</tr>
<tr>
<td>1854 MOTHER/BAP:</td>
</tr>
<tr>
<td>11 APR 1854</td>
</tr>
<tr>
<td>1855 MOTHER/BAP:</td>
</tr>
<tr>
<td>12 AUG 1855</td>
</tr>
<tr>
<td>1857 MOTHER/BAP:</td>
</tr>
<tr>
<td>28 NOV 1857</td>
</tr>
<tr>
<td>1860 MOTHER/BAP:</td>
</tr>
<tr>
<td>11 APR 1860</td>
</tr>
<tr>
<td>1861 CENSUS:</td>
</tr>
<tr>
<td>7 APR 1861</td>
</tr>
<tr>
<td>1862 MOTHER/BAP:</td>
</tr>
<tr>
<td>24 AUG 1862</td>
</tr>
<tr>
<td>1864 MOTHER/BAP:</td>
</tr>
<tr>
<td>10 JUL 1864</td>
</tr>
<tr>
<td>1871 CENSUS:</td>
</tr>
<tr>
<td>2 APR 1871</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EVENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1822 BAPTISM:</td>
</tr>
<tr>
<td>30 SEP 1822</td>
</tr>
<tr>
<td>1849 MARRIAGE:</td>
</tr>
<tr>
<td>21 APR 1849</td>
</tr>
<tr>
<td>1853 MOTHER/BAP:</td>
</tr>
<tr>
<td>4 JAN 1853</td>
</tr>
<tr>
<td>1854 MOTHER/BAP:</td>
</tr>
<tr>
<td>11 APR 1854</td>
</tr>
<tr>
<td>1855 MOTHER/BAP:</td>
</tr>
<tr>
<td>12 AUG 1855</td>
</tr>
<tr>
<td>1857 MOTHER/BAP:</td>
</tr>
<tr>
<td>28 NOV 1857</td>
</tr>
<tr>
<td>1860 MOTHER/BAP:</td>
</tr>
<tr>
<td>11 APR 1860</td>
</tr>
<tr>
<td>1861 CENSUS:</td>
</tr>
<tr>
<td>7 APR 1861</td>
</tr>
<tr>
<td>1862 MOTHER/BAP:</td>
</tr>
<tr>
<td>24 AUG 1862</td>
</tr>
<tr>
<td>1864 MOTHER/BAP:</td>
</tr>
<tr>
<td>10 JUL 1864</td>
</tr>
<tr>
<td>1871 CENSUS:</td>
</tr>
<tr>
<td>2 APR 1871</td>
</tr>
</tbody>
</table>
of 'FRANCES WILSON'. There are details of her baptism and marriage, and also of her presence in the censuses of 1861 and 1871. Finally, her name has been located, as 'mother', in seven baptism records. The details for each event include the date, residence and unique source record identifier, in addition to other relevant information. This collecting together of such a body of information from a number of disparate source records exemplifies the primary goal of nominal record linkage.

Three record types are involved in the recording of life events in the Population Database. These are as follows:

1. P-SOURCE-PRESENT. The sole function of this record type is to ensure that each source record is transferred to the Population Database once only. Since a user is able to control precisely which source records are to be linked in any particular operation it is vital that a careful tally should be kept of the records which have been linked. For example, if there is a marriage record for 'JAMES WATSON' and 'HELEN GREEN', and the user requests that record linkage should be carried out for all records for 'GREEN' then this marriage record will be linked and stored in the Population Database. Should the user subsequently request that records for 'WATSON' are to be linked then it is imperative that the marriage record should not be linked and stored a second time. It is the function of the 'P-SOURCE-PRESENT' CALC record to prevent this happening.
As each source record is linked and stored in the Population Database one 'P-SOURCE-PRESENT' record will be created. The CALC key which is used in this record is the six-character identifier which uniquely identifies the source record, and which was originally created by the Source Loading Subsystem. For example, the key of the marriage record illustrated in Figure 6.2 is 'M00073'. Therefore, before each record is linked an attempt is made to locate the corresponding 'P-SOURCE-PRESENT' record in the Population Database. If it is present then the source record has already been linked, and a further linkage is avoided; if it is not present then the linkage can go ahead.

2. P-CENS-EVENT. This record type is used to register the fact that a particular person has been located in a census household record. When such a household record is linked a 'P-CENS-EVENT' record will be created for each of the occupants, and this will be linked to the appropriate 'P-PERSON' record via the 'P-LIFE-EVENT-SET'. A detailed examination of this process is provided in the next chapter.

A 'P-CENS-EVENT' record contains the relevant personal attribute values which were stored in the original household record. Its main components are as follows:

- source type, e.g. 'C6' (i.e. 1861)
- date of census in internal coded form
- relationship, marital status, date of birth,
birthplace and occupation information in internal coded form.

Other information relating to the encompassing household and enumeration district is stored in separate records. The reasons for this separation and descriptions of the relevant records are provided in Section 9.2.6.

3. P-EVENT. This record type is used to register the fact that a particular person has been referred to in a true or surrogate vital record. When a birth, marriage or death record is linked a 'P-EVENT' record will be created for each of the people mentioned in the record, and this will be linked to the appropriate 'P-PERSON' record via the 'P-LIFE-EVENT-SET' (22).

A 'P-EVENT' record contains the relevant personal attribute values which were stored in the original vital record. Its main components are as follows:

- source type, e.g. 'B1' (i.e. baptism)
- event type, e.g. baptism (father)
- unique source record key, e.g. '800402'
- authority, e.g. 'ANG' (i.e. Anglican)
- date and place event recorded, in internal coded form

information, as is available, about the person owning the 'P-EVENT' record: marital condition, date of birth, occupation and residence, all in internal coded form.
Both 'P-CENS-EVENT' and 'P-EVENT' records are inserted as members in the same set, viz. 'P-LIFE-EVENT-SET'. This is one of only two set types in the complete database which are defined to have two different types of member record. The reason for designing 'P-LIFE-EVENT-SET' in this way was so that for a given occurrence of the set the members, of whichever type, would be ordered in time sequence: this would therefore facilitate the printing of a life-history, as illustrated in Figure 9.6. (23)

9.2.5 The Occupation Index

The primitive population database structure illustrated in Figure 4.11 had a single CALC record, viz. 'OCCUPATION', to provide access to people with a particular occupation. There was also a link record, viz. 'JOB-LINK', to sustain an n:n relationship between the 'OCCUPATION' and 'PERSON' records.

I shall now consider what would be the appropriate characteristics of a more elaborate Occupation Index within the context of the present Population Database. Suitable search criteria are considered to be as follows:

1. Given a particular occupation, e.g. 'SHOEMAKER', locate all those people who had that occupation at some point in their lives.
2. Given a particular occupation, e.g. 'SHOEMAKER', locate all those people who had that occupation in a particular census, e.g. 1851. This more restrictive form of access could facilitate the comparative analysis of different occupational groups at the same census or of the same occupational group at different censuses.

3. For both types of access the index should treat each major occupation name as equivalent to its synonyms. Thus, for example, if the user submits the occupation 'SHOEMAKER' then one would expect the index to be able to search simultaneously on 'CORDWAINER', and on any other occupational synonyms which had been previously designated.

In order to implement this index it is unnecessary for 'JOB-LINK' records to be artificially created, as was the case with the primitive database structure. Arrangements for isolating the occupational information into distinct records separate from the 'P-PERSON' record have already been established in the previous section. The required 'JOB-LINK' records now occur naturally in the form of the event records 'P-CENS-EVENT' and 'P-EVENT'. All that is needed to complete the index, therefore, is to provide appropriate 'OCCUPATION' CALC records for the two types of access described above. These records are as follows:
1. **P-OCC-ENTRY.** This record type is used to provide access to all those people who had a particular occupation at some point in their lives. Within the Population Database there will therefore need to be one 'P-OCC-ENTRY' CALC record for each major occupation name internal code value. Thus, for example, if the internal code value for 'SHOEMAKER' is '0005300' then the corresponding 'P-OCC-ENTRY' record will be stored and retrieved by using the 'CALC' mechanism on the major code value '00053'. For convenience each 'P-OCC-ENTRY' record holds the appropriate character string for the major occupation name. In order to locate all the shoemakers in the database one would first locate the corresponding 'P-OCC-ENTRY' record and then use 'P-OCC-SET' and 'P-LIFE-EVENT-SET' in turn (see Figure 9.5).

Consider first 'P-OCC-SET'. For a given occupation this set will contain a 'P-CENS-EVENT' record for every occurrence of that occupation in the census records which have been linked, and it will contain a 'P-EVENT' record for every occurrence of that occupation in all the other record types that have been linked. Therefore, having located a particular 'P-CENS-EVENT' or 'P-EVENT' record it will be necessary merely to obtain its owner in the 'P-LIFE-EVENT-SET' in order to locate the appropriate 'P-PERSON' record (24).
2. P-OCC-CENS-ENTRY. This record type is used to provide access to all those people who had a specified occupation in a particular census, and there will exist a 'P-OCC-CENS-ENTRY' record for each required occupation-census combination. For example, if there were shoemakers in the 1851 census then there will exist a shoemaker-1851 'P-OCC-CENS-ENTRY' record. To locate this record one would use the 'CALC' mechanism on the major code for 'SHOEMAKER', e.g. '00053', and 'C5', i.e. on the combined code '00053C5'.

Once a particular 'P-OCC-CENS-ENTRY' record has been located access to the individual 'P-PERSON' records will be achieved via the 'P-OCC-CENS-SET' and the 'P-LIFE-EVENT-SET'. For a given occupation the 'P-OCC-CENS-SET' will contain a 'P-CENS-EVENT' record for every occurrence of that occupation in the census records for the corresponding census. From each 'P-CENS-EVENT' record it will be necessary merely to obtain its owner in the 'P-LIFE-EVENT-SET' in order to locate the required 'P-PERSON' record.

9.2.6 The Place Index

The primitive population database structure illustrated in Figure 4.11 had a single CALC record, viz. 'LOCATION', to provide access to people who had lived in a particular place. There was also a link record, viz. 'PLACE-LINK', to sustain an n:n relationship between the
'LOCATION' and 'PERSON' records.

It was recognised at an early stage that the provision of a fully generalised Place Index would not be within the scope of the present research. For the reasons given in Section 3.2.1 the handling of spatial information is complicated, in particular because of the problems of variable precision and changing boundaries. As a partial solution, however, it was decided to establish in the Population Database information about census enumeration district and household composition and provide an associated index.

What this scheme provides is the ability to select a particular enumeration district and a particular census and locate all the relevant households and people present. Using this type of access one could proceed, for example, to carry out a comparative analysis of two different communities by studying the populations present at the same census. Alternatively, one could select just one community and study its changing characteristics over time by observing the populations present at two or more censuses (25).

The information about enumeration districts and households is stored in separate records. These are as follows:

1. P-CENS-ENTRY. This record type is used to provide access to a particular enumeration district at a particular census, and there will exist a 'P-CENS-ENTRY' record for each required enumeration district-census combination. For example, there
will be a record for Elwick Hall Parish-1851. To locate this record one would use the 'CALC' mechanism on the major code for 'ELWICK HALL PARISH', e.g. '00072', and 'C5', i.e. on the combined code '00072C5'.

Once a particular 'P-CENS-ENTRY' record has been located access to individual households, and subsequently to their occupants, may be achieved by first accessing the 'P-CENS-HOUSEHOLD' records in the 'P-CENS-HSE-SET': there will exist one such record for each household in the enumeration district for a particular census.

2. P-CENS-HOUSEHOLD. Each 'P-CENS-HOUSEHOLD' record corresponds with an 'S-CENS-HOUSEHOLD' record in the Source Database (26). The main components of a 'P-CENS-HOUSEHOLD' record are therefore as follows:

   unique source record key, e.g. 'C00051'
   household address information
   counters, indicating how many people there are in the household, how many of them are related to the head and how many have the same surname as the head.

Once a particular 'P-CENS-HOUSEHOLD' record has been located access to the individual 'P-PERSON' records may be achieved via the 'P-CENS-OCC-SET' and 'P-LIFE-EVENT-SET'. For a given household the 'P-CENS-OCC-SET' will contain a
'P-CENS-EVENT' record for each person in the household: the corresponding 'P-PERSON' record is obtained by finding the owner in the appropriate 'P-LIFE-EVENT-SET'.

I have now completed my analysis of the major structural components of the Population Database. In the next chapter I shall explore the nature of the strategies which are needed to transfer source records to the Population Database and link them.
NOTES

1. Strictly, the calling sequence is as indicated in Figure 9.4, i.e. that 'PZDBPROCESS' calls 'DTCONTRACT' and that 'DTCONTRACT' then calls 'DUSNSTREAM'. It is necessary for all calls of 'DUSNSTREAM' and 'DVDBSEARCH' to come via 'DTCONTRACT' because of IDMS 'BIND' constraints. In short, the database record buffers which 'DUSNSTREAM' and 'DVDBSEARCH' both need to use are declared in the WORKING-STORAGE SECTION of 'DTCONTRACT', and they can only be made available to these modules if they are passed as parameters in a call from 'DTCONTRACT'.

2. The calling sequence is via 'DTCONTRACT'. See Note 1.

3. Since a woman can, in principle, have many married surnames it is inappropriate to attempt to store them in the 'P-PERSON' record. As will be shown later each such surname occurrence is more conveniently associated with the relevant 'P-FAMILY' record.

4. Such birth information can be obtained or deduced from more than one record, e.g. from successive census records, and so the values held in the 'P-PERSON' record should ideally represent an 'average' of all the occurrences encountered. For example, for an age the value held could be the arithmetic mean, while for a birthplace it could be the most commonly occurring placename. In the present implementation the values used are those obtained from the first occurrence encountered. All other occurrences are, however, not lost: they are stored, along with other associated information, in the various 'P-EVENT' and 'P-CENS-EVENT' records.

5. Such a status value is, however, not normally provided in the type of burial records (i.e. post-1812) which were used.

6. These counter values are, like the linkage status indicators, strictly unnecessary, since they can be obtained directly by counting the number of member records in the three sets referred to. In effect, the counters represent a duplication of information, which in database design terminology is referred to as 'data redundancy'. Such redundancy is normally introduced for reasons of efficiency and/or convenience. However, in view of the possibility of a discrepancy arising between the two manifestations of the same information (for example, through a programming error) the use of data redundancy should be approached with care. See Martin 1975, 32-3 and 509-10.

7. The holding of such information in the 'P-FAMILY' record constitutes another example of data redundancy. See Note 6.

8. By contrast, in an implementation using a relational database system the actual connection would need to be effected by program statements and the explicit manipulation of the key values which
are used to identify records. It is clear, therefore, that in this respect the CODASYL approach provides the more appropriate mechanisms for making the inter-record linkages.

9. In the case of persons and families it would appear that two parallel schemes are being used to count the same item: one via the 'P-ITEM-CNT' record used to furnish unique identifiers, and the other via the 'P-ITEM-CNT' counter records. However, it is necessary to maintain these separate schemes because of the situation which arises where a particular record linkage operation requires two 'P-PERSON' records, say, to be merged. This could occur, for example, where a 'P-PERSON' record had been created for an unmarried female located in the 1851 census and another had been created for a married female located in the 1861 census. Should it subsequently be discovered from a marriage record that the two 'P-PERSON' records represent in fact the same individual then it will be necessary to merge the information into one of the 'P-PERSON' records and destroy the other one. And at the same time one would need to reduce by one the count of the number of people in the database. Clearly if the two 'P-ITEM-CNT' records which are used to furnish unique identifiers for people and families were to be decremented then ambiguities could arise. The maintenance of additional counter records is therefore essential.

10. Ideally, in a more generalised implementation one would not wish to decide in advance which subgroups are likely to be 'of potential interest': rather, the user should be able to nominate in an ad hoc fashion the particular subgroups in which he is interested. A scheme which would enable the present facilities to be extended in this way is discussed in Section 11.4.

11. For the purposes of this selection any person who is recorded as being born in a 'peripheral' location is treated as being a native of the base zone. In a more sophisticated implementation one would allow the user to decide dynamically how such persons should be grouped.

12. It is to be expected that as record linkage proceeds and more information is accumulated it may become possible to 'promote' people from this fifth subgroup into one of the other four. However, in view of the potentially large membership of the subgroup it might be found desirable to subdivide it into a number of smaller groupings: for example, subdivision by gender would probably constitute a worthwhile design improvement.

13. Rather than create the 'P-PERSON-COHORT' records on an ad hoc basis as they are needed it was decided that a potentially more efficient strategy would be to create the more frequently needed records at the time when the Population Database is set up. This policy has been adopted, and for each of the four major subgroups 'P-PERSON-COHORT' records are created for the years 1771-1871. Should it subsequently be required to enter into the database a 'P-PERSON' record for someone who was born before 1771 then it would be necessary to create dynamically a 'P-PERSON-COHORT' record.
record for the appropriate year, gender and birthplace code.

14. This third subgroup provides a 'catch-all' category similar to the one provided for 'P-PERSON' records. And in a similar fashion as record linkage proceeds and more information is accumulated it should be possible to 'promote' families from this subgroup to one of the other two. Likewise, it would be possible to subdivide the category if its membership was found to be unduly large.

15. The policy for creating the corresponding 'P-FAMILY-COHORT' records follows the one adopted for the 'P-PERSON-COHORT' records (see Note 13). Thus, at the time when the Population Database is created, 'P-FAMILY-COHORT' records are created for the two major subgroups for the years 1771-1871.

16. For each female the global access to the 'P-PERSON' record should be via both maiden and married surnames. Thus, for example, for the surname 'SMITH' there should be access to all the 'P-PERSON' records for all females for whom this is the maiden or married surname.

17. The inclusion of this character string can facilitate the display of information held in the Population Database, since it obviates the need to access the Directory Database. Such character strings are also held in the other records in the Name Index, and they are also used in the Occupation and Place Indexes.

18. This particular example serves to demonstrate that a link record need not exist merely as an artificial entity to support an n:n relationship. In the present example the link record, viz. 'P-FAMILY', has an obvious and very important 'real world' significance.

19. This assumption may not always be valid since it is possible for an individual to be posthumously referred to in a record, without an appropriate acknowledgement to this effect.

20. The exploratory discussion about population databases in Chapter 4 dealt with the problems of storing information relating to a person's occupation and residence, but it did not attempt to extend the concepts to provide a generalised life event system, of the kind described here.

21. With the system which is currently operational marriage events are created only for the marriage partners and not also for the fathers. This simplification was adopted chiefly because only the post-1837 marriage records contain the additional information about fathers.

22. The chief justification for maintaining distinct kinds of records for census events and vital events is that it is potentially valuable when recording census events in the Population Database to retain information about the encompassing household and enumeration district structures. Such structures require the
provision of additional records and set connections, not needed when recording vital events.

23. Unfortunately, because of an error or restriction in IDMS the records in this set were not maintained in the required order: all the 'P-CENS-EVENT' records came first in time order, followed by all the 'P-EVENT' records in time order. It was necessary, therefore, to provide some additional program code to merge the two sequences when a life-history was to be displayed. Happily, this was the only occasion when IDMS did not perform in strict accordance with its specification.

24. Since it is possible for a given 'P-OCC-ENTRY' record to be linked to a particular 'P-PERSON' record via several 'P-EVENT' and/or several 'P-CENS-EVENT' records a traversal of the 'P-OCC-SET' and 'P-LIFE-EVENT-SET' could lead to the same 'P-PERSON' record several times over. For example, if a person had been described as a 'SHOEMAKER' in three source records then the Occupation Index would furnish three routes to his 'P-PERSON' record via the corresponding event records. As a result, if asked to locate each shoemaker then the index might inadvertently identify the same person more than once. To avoid such potential confusion a special marker value is placed in one event record for each person for a given occupation: this serves to distinguish a single access route between any pair of 'P-OCC-ENTRY' and 'P-PERSON' records.

25. With the present implementation, however, the Population Analysis Subsystem does not support the appropriate commands to facilitate such analyses.

26. There is, therefore, no attempt to link physical houses to households across the three censuses. (For an elaboration of some of the issues involved in such linkage see Henstock 1973.) If such linkage were to be carried out then appropriate modifications would need to be made to the Place Index. In particular, the 'P-CENS-HOUSEHOLD' record would need to be replaced by a 'P-HOUSE' record. Each 'P-HOUSE' record would then need to become the owner in a 'life-event' set and own 'event' records in a way which would parallel the scheme provided for 'P-PERSON' records.
In the last chapter I demonstrated how individual records from the Source Database could be selected and displayed, and I also analysed in detail the major structural components of the Population Database. By this stage I have therefore established quite firmly the nature of the transformations which must occur as nominal records are transferred from the Source Database to the Population Database. The significant task which remains is to examine the nature of the transformation processes themselves, i.e. how nominal records can actually be restructured and linked in the Population Database. This will be the focus of the present chapter.

I shall begin by examining the required functional properties of the linkage strategy in relation to the characteristics of the particular kinds of nineteenth century nominal records which are to be linked. I shall then outline the main operational characteristics of my proposed strategy. The bulk of the remainder of the chapter will explore in detail the way in which the strategy is applied in the case of each of the four main kinds of nominal data: census household records, marriages, baptisms and burials. Next I shall examine the way in which the linkage processes are implemented, by looking at the internal design of the Record Linkage Subsystem. I shall then assess the validity of the linkage strategy in the light of evidence from nineteenth century census data concerning the extent of nominal
ambiguity. Finally, at the end of the chapter, I shall provide a critical re-appraisal of the entire strategy and make some proposals about possible, future improvements.

10.1 RECORD LINKAGE

10.1.1 The Overall Linkage Strategy

In my analysis of the problems of nominal record linkage in Chapter 3 it was observed that each historical context presents its own distinctive set of characteristics and associated problems. Nominal ambiguity was seen to be a serious problem when attempting to link seventeenth century records; while for nineteenth century records it was suggested that the organisation of family-based linkage and the handling of structurally complex data were the most significant problems.

I shall begin by considering the linkage of census household records for the years 1851, '61 and '71. The task of carrying out person-based linkage for individuals between one census and another is technically no more complicated than for other types of nominal records, and the presence of information about an individual's age, birthplace and marital status can significantly reduce the problems of nominal ambiguity, particularly for small communities. The linkage of
census data, however, becomes much more complicated when we move beyond the constraints of person-based linkage and attempt to treat the unit of linkage as the co-residing family, rather than the individual household occupant. Because we are required to match and link such a wide range of disparate family groupings as are presented in census households the linkage task is conceptually and organisationally very complicated.

The linkage of nineteenth century parish register records presents the more familiar nominal record linkage problems. Thus, while the records for one type of event, e.g. baptisms, possess structural homogeneity they do not contain a significant amount of information which can aid the matching process. Let us consider the properties of each type of record in turn, as follows:

1. Baptisms. For the person baptised the date of baptism can provide, in general, an approximate date of birth. Likewise, the place of baptism will also usually provide an approximation to the place of birth. However, in some cases (e.g. for adult baptisms) this method of deducing the date and place of birth will be invalid, and therefore the technique needs to be applied with due caution. The names of the individual's parents can also assist the linkage process, particularly where such information is contained in other records (e.g. census) which are being linked.
When we link a baptism record for a legitimate birth we will normally be attempting to carry out nominal record linkage operations for three people: the father, the mother and the child. Unfortunately for the father and the mother only minimal information is provided: in particular, there is no indication of date of birth and place of birth. The only 'clue' to their dates of birth is provided by the date of birth of the child. And for the mother there is frequently no indication of her maiden name. Linkage for such individuals must therefore be a tentative operation.

2. Marriages. Post-1837 records are relatively simple to link since, in general, they possess for each of the marriage partners the name (including the maiden name for the bride), age information, marital status and the name of the father. Pre-1837 records usually lack the age information and name of the father, and so are roughly as difficult to link as the parents in a baptism record.

3. Burials. Post-1812 records are moderately simple to link since they contain the age of the deceased. However, the absence of marital status information can cause problems when linking records for women, and the absence of information about relatives (e.g. parents or spouse) will generally increase the problems of correct identification.
It should be clear that the range of nineteenth century nominal records for which linkage strategies are to be provided is a wide one: from the structurally complex and detailed post-1841 census household record to the simpler and more imprecise baptism record. In addition, it should be noted that a family-based linkage strategy will need to be capable of establishing linkages of two distinct kinds. Not only will it be required to link a particular person to all the records in which he is named, but also it will be required to link each person to his parents, spouse(s) and children.

I shall now examine the practical issues which must be addressed when designing and implementing a linkage strategy. The overall design of the strategy must clearly be related to the types of records which are being used and to their information content. Where the records have a relatively low information content then there are likely to be serious matching problems. For such records it is therefore imperative that, as far as possible, linkage decisions should be based on all the relevant information which is available. Thus, Wrigley and Schofield include in their proposals for linking English parish register records the recommendation that 'the decision procedure should make the best use not simply of the records immediately involved but of all the information in the file as a whole which is relevant to the decision' (Wrigley and Schofield 1973, 68). The same overall policy is implicit in strategies for linking Italian parish register records of mediocre quality (Skolnick 1973).
In general, linkage strategies consist of two basic steps: these have been described as 'a searching step' and 'a detailed comparison step' (Winchester 1970, 113). During the searching step the objective is to select from all the records available those which are potential candidates for a particular linkage operation. And then during the detailed comparison step the objective is to decide which of the candidate records (if any) should be linked.

Let us examine the searching step in more detail. A common approach, much used by researchers in the field of medical record linkage, is to group records into sorting 'pockets' on the basis of the Soundex code derived from the individual's surname (Winchester 1970, 114-7). The use of the 'N-SURNUM' record in the Source Database implementation (see Figure 8.6) serves a similar function. For more complex linkage situations, for example, where parish register records are to be linked, the process is rather more complicated, normally involving the creation of record 'clusters' (Skolnick 1973; Wrigley and Schofield 1973). Each cluster consists of a group of baptism, marriage and burial records which are potential candidates for linkage.

Various techniques are used during the detailed comparison step to establish which of the candidate records should be linked. These are based on matching strategies in which there is a field-by-field comparison of the records. Thus, for example, in a particular linkage situation one might be comparing in two records the fields containing the surname, Christian name, age and birthplace, the aim being to
establish whether there is a sufficient match between these items to confirm that the records should be linked. At the simplest level one can, for example, adopt a matching strategy in which one scores +1 for each pair of identifying items which are in agreement, -1 for each pair which are in disagreement and 0 for the situation where there is missing information. By adding the individual scores one obtains a total score which in some way indicates the strength of the link. Alternative, more elaborate strategies are also available (Winchester 1970, 117-22).

The more complicated linkage strategies are often designed to operate in an iterative fashion. Thus, for example, the scheme proposed by Wrigley and Schofield employs ten distinct stages (Wrigley and Schofield 1973, 69-70), most of which are concerned with a specific type of inter-record linkage. Stage 4, for example, is concerned with the linkage of marriage partners to their baptism records. At the beginning of the whole operation all the possible links which might exist between all records are set up. Subsequently, by successive iteration, impossible and unlikely links are deleted until the set of linkages which remain are deemed to be the most likely and mutually compatible ones. Skolnick, basing his strategies on the techniques of artificial intelligence, also attempts to achieve a solution by a process of successive iteration and refinement (Skolnick 1973, 122-5). In his case he creates a 'family of related solutions', from which he progressively eliminates the most unlikely, and at the end of the process the aim is to be left with the 'best solution'. But problems have been observed with the use of such
iterative techniques (Macfarlane et al 1983, 29). In the first place they can consume very large amounts of computer time. In addition, some of the computations which are involved are 'ad hoc and impressionistic' and so the 'correctness' of the results is in doubt.

There are two factors, in particular, which militate against the use of such iterative strategies in the context of nineteenth century demographic data. In the first place it is difficult to see how they can be generalised to encompass the complex and arbitrarily structured census household record. One of the fundamentals of their central logic is that all records of a particular type have the same essential structure: thus, one marriage record is similar to all other marriage records. Unfortunately, the same is not true for census household records. Each household contains an arbitrary number of arbitrary pieces of genealogical structure, and in the process of linkage these must be linked to other such pieces, as well as to other types of records. The basic regularity of parish register records facilitates the creation of sets of rules and procedures for linking them; the irregularity of census household records makes the creation of corresponding rules and procedures appropriate to them much more difficult.

The second factor which militates against the use of such iterative strategies for nineteenth century data concerns the information content of records. As was pointed out earlier there is a whole range of record types, from the highly detailed census household record to the simpler burial record. In these circumstances it was
concluded that the resolution of nominal ambiguity would not be the central problem, at least when dealing with smaller communities (1). Instead, the complexity of handling the range of information provided, and the associated problems of assembling it into genealogical structures, were considered to present the more significant areas of difficulty.

Taking these factors into consideration it was decided that a quite different approach to the record linkage task was needed. In its basic operation the new strategy would need to be sequential rather than iterative, and it would function in a way which paralleled the manual procedures of the genealogist. In terms of the concepts of Source and Population Databases, as already described, the main operational characteristics of the strategy which has been devised are as follows:

- The record linkage is to proceed in a sequential fashion. This involves taking the records, one at a time, from the Source Database and transferring them to the Population Database. As each record is transferred an attempt is made to locate in the Population Database each person referred to in the record. For those so located the information from the record is added to the information already held; for anyone who is not located a new 'person' is created and initial information associated with them. Simultaneously, appropriate familial connections are also set up between the people held in the database. Record linkage therefore operates in a
cumulative fashion, and at any moment the contents of the Population Database represent the results of the linkage thus far and the body of information against which the contents of the next record to be linked will be matched.

- The most 'information-rich' records are to be linked first. The default order in which records are presented for linkage is as follows:
  census
  marriage
  birth/baptism (2)
  death/burial (3)
This strategy ensures that the 'easiest' records (i.e. census) are linked first, and that when the time comes to link the more 'difficult' baptism and burial records there will exist in the Population Database a large body of information against which to match these records (4).

- The user is to be allowed to select the order in which the records are linked, where he wishes to override the ordering defined above (5). Facilities which enable him to do this were described in Section 6.5.

- The 'correctness' of the linkage is to be measured by analysing the final contents of the Population Database. The underlying assumption in this approach is that with the kinds of data being used the strategies described should, in the
main, be able to establish unambiguous, 'correct' linkages. In order to confirm that this is, indeed, the case it should be necessary merely to subject the results of the linkage operations to appropriate post-linkage analysis (6).

Two separate ways of carrying out this analysis are envisaged. The most obvious way is to take some sample records, say for a particular name, and link them manually. The results can then be compared with those produced by the system. The second, and perhaps more convenient, way would be to attempt to assess the 'correctness' of the linkage by carrying out some statistical analyses of the contents of the entire Population Database. For example, if there has been a tendency to 'over-link' census records, i.e. make connections which should not be made, then we might expect this to be particularly apparent for people with popular names, simply because there will be a larger number of candidate records available for linkage. How then should one attempt to detect the presence of over-linkage? One possible way would be to observe the discrepancies between the computed dates of birth derived from the census records for each individual. On average we would expect true links to display smaller computed date of birth discrepancies than false links. If therefore we were to find that people with popular names display greater computed date of birth discrepancies than those with less popular names we would suspect that 'over-linking' had occurred. I shall explore the use of this technique in
Before analysing the linkage strategy in detail I shall make a number of observations about the overall programming strategy which was considered to be most appropriate.

10.1.2 Programming Strategy

For any particular programming problem there are usually several types of solution available, and it is crucial that the selection should be made carefully. A sharp distinction can be drawn between the following two major problem-solving strategies:

1. the algorithmic approach. In this case the individual steps in the solution of the problem are contained within the sequence of statements in the program itself. And if it is required to change the algorithm then the program logic must be changed.

2. the table-driven approach. In this case the information which provides the solution to the problem is maintained in one or more entries in one or more tables, and an associated program is required only to manipulate the table entries. Where it is required to modify the solution strategy changes are made to the table(s), but not to the program.
In general, it can be said that the algorithmic approach is normally simpler to implement than the table-driven approach, but also that it is less flexible. Let us consider this in relation to the two alternative methods for handling name variations which I described in Section 3.1.2. Firstly, the Russell-Soundex coding algorithm. This is a moderately simple program to write, but it invariably has to be tailored to the idiosyncrasies of the 'name universe' which it is required to handle. Where such an algorithm is to be used to handle a different name universe then it will be necessary to analyse the characteristics of this new name universe and make appropriate amendments to the algorithm (Blayo 1973, 57-8; Wrigley and Schofield 1973, 98-101). By contrast, the use of name directories, which is essentially a table-driven approach, is more difficult to set up, but modifications to the name universe can be accommodated without any reprogramming.

Clearly, when seeking a strategy for accomplishing nominal record linkage one would naturally choose to adopt the more flexible table-driven approach. One might then envisage supplying the system with two sets of tables, one describing the characteristics of the nominal record types which are to be linked, and the other the characteristics of the linkage strategies to be applied. Subsequently, and without any reprogramming, one could enhance the system to handle new types of nominal records and/or to apply modified linkage strategies.
Despite the obvious attractions of such a table-driven approach there can also be significant disadvantages. As a technique it is invariably more difficult to implement than the algorithmic approach. And considering the use of the technique within the specific context of nominal record linkage it would seem to imply that my design objective was to produce a generalised rather than a special-purpose solution. But, as observed in Chapter 3, such a goal seems to be unrealistic, given the extent and seriousness of the linkage problems associated with a generalised solution.

It was with these considerations in mind that the decision to adopt the simpler, algorithmic approach for the record linkage strategy was taken. In the following sections I shall explore the nature of the resulting linkage strategy, and shall observe the mechanisms which are invoked when each of the four main types of records is presented for linkage. At the end of the chapter, in Section 10.3, I shall return to the question of programming strategy and shall make some proposals about the kinds of steps which would lead to a more generalised approach.

10.1.3 The Initialisation of the Population Database

Before examining specific linkage operations I shall consider the nature of the processes which are involved in initialising the Population Database and making it ready. Such an initialisation will be carried out in response to the user command 'PP/DBINIT', as was
described in Section 6.5.

Let us consider what needs to take place during initialisation. Essentially the task involves the creation and initialisation of a number of records. The required actions are as follows:

- Firstly a check must be made to ensure that the Population Database is a new one. The technique adopted is the same as that used during the initialisation of the Directory and Source Databases, i.e. the presence of a special 'marker' record is checked. If it already exists then the initialisation is curtailed and an error message is sent to the user.

- The special 'P-ITEM-CNT' marker record is created.

- For both persons and families a 'P-ITEM-CNT' record is created containing the initial value '1'. These are the records which are subsequently used to furnish the unique six-character identifiers for newly created 'P-PERSON' and 'P-FAMILY' records.

- The five additional 'P-ITEM-CNT' records which will be used to count persons, families, surnames, person names and family names are created and given initial values of '0'. 
The final action is to initialise the Cohort Index by creating an appropriate number of the relevant index records: 'P-ALL-PERSONS', 'P-PERSON-COHORT', 'P-ALL-FAMILIES' and 'P-FAMILY-COHORT'. Details of this initialisation were provided in Section 9.2.2.

10.1.4 Linking Census Household Records

In this and in each of three subsequent sections I shall take a particular record type and shall examine the problems of transferring it to the Population Database and linking it. For each type of record I shall be concerned with the way in which the linkage strategy needs to be adapted to suit the particular characteristics of the record. At the end of the chapter, in Section 10.3.2, I shall examine a number of refinements which might lead to a more unified strategy, and one which is more source-type independent than the present one.

In this section I shall consider the linkage of census household data, and I shall do this by examining how a particular census household record, having been selected from the Source Database, is transferred to the Population Database and linked.

There are two fundamental, but quite distinct, issues involved in the process of linking a record, and these are particularly well exemplified in the case of census household data. The first issue is concerned solely with information transfer. In the process of linking
a census household record the information from the record must be structurally transformed as it passes to the Population Database. During the transfer the essentially dominant 'person-within-household' grouping of the original source record needs to be made subordinate to the more important and fundamental 'person-within-family' grouping. In order to achieve this transformation the family relationships which are implicit in the census household structure need to be extracted and made explicit. This operation, which is an integral element of family-based record-linkage, can usefully be viewed as a particular kind of 'linkage' in its own right. Thus, while in the original source record the linkage between people is primarily a residential one, when the structural transformation takes place it becomes necessary to establish new, familial linkages between them (7).

Given that it may be possible to make the appropriate structural transformations to the information in a census household record one is then confronted with the second fundamental issue, viz. how is the restructured information to be linked to information already present in the Population Database? And more especially, given a particular individual in a household record, how is one to decide whether the person is already present in the Population Database or whether a new record needs to be created for him? This decision will clearly demand that a matching strategy should be invoked, to make comparisons between information in the household record and information in the Population Database. When the decision has been made the information can then be inserted in the database in an appropriate fashion, and with connections made, as necessary, to the information already held.
Let us begin then to look in a more detailed way at the complete strategy. But let us start by considering in basic terms how information is transferred to the Population Database, and let us imagine the simplest possible situation: the transfer of the first record to the Population Database, and where this record consists of a household with only one male occupant. Such an operation will require no record linkage, but only information transfer.

The relevant information from the Source Database will be contained in one 'S-CENS-HOUSEHOLD' record and one 'S-CENS-PERSON' record. When transferred to the Population Database it will need to be set up in three connected records: a 'P-CENS-HOUSEHOLD' record, a 'P-CENS-EVENT' record and a 'P-PERSON' record (see Figure 9.5). Essentially this will represent a person for whom there is one life-event, and simultaneously a household in which there is one occupant. Having done this it will then be necessary to connect these records to appropriate index records. The 'P-PERSON' record will need to be connected to the Cohort Index via the appropriate 'P-PERSON-COHORT' record. In a similar fashion the 'P-PERSON' record must be connected to the Name Index: in this case, since this will be the first person to be connected to the index it will be necessary for the required 'P-SURNAME' and 'P-PERSON-NAME' records to be created. If an occupation has been specified then the 'P-CENS-EVENT' record will need to be connected to the Occupation Index: this will require the creation of the necessary 'P-OCC-ENTRY' and 'P-OCC-CENS-ENTRY' records. Finally, the household must be connected to the Place Index: this will involve the creation of a 'P-CENS-ENTRY' record for the
I shall now repeat the above analysis, but this time I shall consider what additional operations must take place when more typical household records are involved and when linkage is needed. The first action which must take place when transferring a census household is, as before, for a 'P-CENS-HOUSEHOLD' record to be created in the Population Database, and for information about the household to be stored in it. The record will be connected to its appropriate 'P-CENS-ENTRY' record in the 'P-CENS-HSE-SET': if this is the first record in the .set then it will also be necessary to create the 'P-CENS-ENTRY' record. As a result of these operations information about the household will be stored in the database and will be connected to the information about the associated enumeration district.

The majority of census household records contain details of a husband and wife, together normally with details of various relatives and other individuals. An important first step in the linkage process, therefore, is to try to establish whether this family unit has already been registered in the Population Database, e.g. from a previous census. Clearly it should be possible to match several people in combination more confidently than to match them in isolation (Hershberg et al 1976A, 140; Bouchard and Pouyez 1980, 123). It would, of course, be possible to include additional members of the family (e.g. children) when matching family units, although it is probably not justified in view of the additional complexity involved.
Therefore the first step in the present strategy is to attempt to find a match for the husband and wife together. If the husband and wife combination has already been registered then there will exist a 'P-FAMILY' record for them, and this will be linked to the corresponding 'P-FAMILY-NAME' record. An attempt is therefore made, using the appropriate husband and wife name codes, to locate the corresponding 'P-FAMILY-NAME' record. If it does not exist then their 'P-FAMILY' record also will not exist, and so the search can be terminated. If it does exist then the implication is that there is at least one, and possibly several, candidate 'P-FAMILY' records, i.e. records which have matching major husband and wife name codes. These records, which will exist as members of the corresponding 'P-COUPLE-NAM-SET', can therefore be interrogated in turn until one is found whose details closely match those of the husband and wife in the census household record (8). The present strategy employs relatively simple criteria for deciding whether a matching 'P-FAMILY' record has been located. Thus, a 'P-FAMILY' record is deemed to match if either:

1. there is no discrepancy in the birthplace codes for both husband and wife and also the discrepancies in the computed dates of birth for both husband and wife do not exceed six years or

2. there is no discrepancy in the birthplace codes for one of the marriage partners and also the discrepancies in the computed dates of birth for both husband and wife do not exceed two years (9).
If an acceptable 'P-FAMILY' record can be located in this way then it may be used as the main 'reference' family for all the subsequent linkage operations for the current household. This family will subsequently be referred to as the 'primary' family unit. What it essentially gives is access to the accumulated family information in the Population Database against which information from the household record can be matched.

After this initial attempt to locate the primary family unit the remaining step in the linkage process is to take each household occupant in turn, check the flag which indicates whether the occupant is 'linkable', and, if he or she is, then carry out the necessary linkage operations.

In order to simplify the following description as much as possible I shall proceed initially by considering the linkage of an occupant who is recorded as being not related to the head of the household. For such a person one is able to engage only in person-based linkage operations.

The first task is to determine whether the person already exists in the Population Database. The search for the person proceeds in approximately the same way as the corresponding search for a family, as described above. Thus, if the person is present in the Population Database then there will exist a 'P-PERSON' record for them, and this will be linked to the corresponding 'P-PERSON-NAME' record in the Name Index. But it should be noted that if the household occupant is a
married or widowed female then the link to the appropriate 'P-PERSON-NAME' record will not be directly via the 'P-BIRTH-NAME-SET', but instead via the 'P-MOTHER-SET', a 'P-FAMILY' record and the 'P-WIFE-NAME-SET'. The first step is to use the occupant's major forename and surname codes to locate the corresponding 'P-PERSON-NAME' record. If it does not exist then their 'P-PERSON' record also does not exist, and so the search can be terminated.

If the 'P-PERSON-NAME' record does exist then the search can continue. If the household occupant is male, or female and unmarried, then the 'P-PERSON' records which are members of the corresponding 'P-BIRTH-NAME-SET' are interrogated in turn until one is found whose details closely match those of the occupant (10). If the household occupant is a married or widowed female then the 'P-FAMILY' records which are members of the 'P-WIFE-NAME-SET' are located in turn, and for each of these the corresponding 'P-PERSON' record which is its owner in the 'P-MOTHER-SET' is located. The interrogation of these 'P-PERSON' records proceeds until one is found whose details closely match those of the occupant.

As for the matching of 'P-FAMILY' records the corresponding criteria for matching 'P-PERSON' records are relatively simple. The criteria are as follows:

- if the household occupant record is, say, from the 1851 census then an initial check is made to determine whether the
'P-PERSON' record under consideration has already been located in the 1851 census. If it has then 'no match' is recorded; if it has not then the matching continues as follows.

- The corresponding fields in the household occupant and 'P-PERSON' records are compared and a simple matchscoring algorithm is executed. In essence this type of scheme provides a convenient means for measuring the amount of agreement between parallel information fields in two records. For each condition listed below there is an associated score. In the comparison of the fields in two records a total score is built up by adding together the individual scores for those conditions which are found to be true. If the total score exceeds some particular, preset 'cut-off' value then the records are deemed to match (11). The conditions and associated scores are as follows:

<table>
<thead>
<tr>
<th>Condition</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>matching gender</td>
<td>1000</td>
</tr>
<tr>
<td>identical forename</td>
<td>10</td>
</tr>
<tr>
<td>identical surname</td>
<td>10</td>
</tr>
<tr>
<td>matching second initial</td>
<td>10</td>
</tr>
<tr>
<td>second initial not present</td>
<td>5</td>
</tr>
<tr>
<td>computed date of birth comparison</td>
<td></td>
</tr>
<tr>
<td>either date not present</td>
<td>100</td>
</tr>
<tr>
<td>agreement (within precision)</td>
<td>200</td>
</tr>
<tr>
<td>within 2 years</td>
<td>160</td>
</tr>
</tbody>
</table>
2-4 years discrepancy

birthplace code:
- either code not present
- both codes are 'native' or 'peripheral'
- one code is 'non-native' and the other is 'non-native' or 'peripheral'

birthplace:
- either placename not present
- identical placenames
- identical major placenames

- if the total score obtained in any matchscoring operation exceeds the 'cut-off' value 2100 then the household occupant and 'P-PERSON' records are deemed to constitute an acceptable match. Although the various components of this algorithm were assembled in a fairly arbitrary way, it has performed satisfactorily without the need for adjustment. It is recognised, however, that some 'fine-tuning' would probably be necessary if the system were to be used with a larger dataset.

If the search for the household occupant in the Population Database proves to be unsuccessful then it will be necessary to create a new 'P-PERSON' record for the occupant. This will involve the
transference of the basic, identifying information from the census record to the 'P-PERSON' record and then the connection of this record to the Cohort and Name Indexes. The detailed operations are as follows:

- initialise various fields and indicators in 'P-PERSON'.

- obtain a unique person identifier and store it in 'P-PERSON'.

- store also name, gender, computed date of birth and birthplace details.

- transfer the 'P-PERSON' record to the Population Database, i.e. 'create' it.

- establish a computed year of birth for the household occupant (i.e. from the computed date of birth), locate the appropriate 'P-PERSON-COHORT' record, and connect the 'P-PERSON' record in the 'P-PERSON-SET'.

- if the household occupant is male, or female and unmarried, then use the major forename and surname codes to locate the corresponding 'P-PERSON-NAME' record, or to create it if it does not exist, and connect the 'P-PERSON' record in the associated 'P-BIRTH-NAME-SET'. 
- if the household occupant is a married or widowed female then it is necessary to connect the 'P-PERSON' record to the Name Index via the 'P-WIFE-NAME-SET', the 'P-MOTHER-SET' and a specially created 'P-FAMILY' record. The appropriate 'P-PERSON-NAME' record is located in the same way as for a male occupant, i.e. by using the major forename and surname codes. The arrangements for setting up the special 'P-FAMILY' record are rather similar to those for setting up the 'P-PERSON' record, although in this case there is only a minimal amount of information to be stored.

- increment by 1 the 'P-ITEM-CNT' record which is used to count the number of persons in the Population Database.

Having located or else created the 'P-PERSON' record for the household occupant all that remains is to create a 'P-CENS-EVENT' record for them and connect it in their 'P-LIFE-EVENT-SET'. This will register the fact that they have been identified in a census household record. The necessary operations are as follows:

- transfer various fields from the 'S-CENS-PERSON' source record to 'P-CENS-EVENT'. These include the relationship, computed date of birth, occupation and birthplace information.

- where the 'P-PERSON' record lacks a particular significant item of information, such as the computed date of birth, copy this from the 'P-CENS-EVENT' record (13).
- set the appropriate indicator in the 'P-PERSON' record to register the particular census in which the person has been located.

- transfer the 'P-CENS-EVENT' record to the Population Database, i.e. 'create' it. IDMS will automatically arrange to connect the record in the appropriate 'P-LIFE-EVENT-SET' and 'P-CENS-OCC-SET', i.e. it will connect it to the appropriate person and household.

- finally, if the 'P-CENS-EVENT' record contains an occupation value then it will be necessary to connect the record to the Occupation Index. This will require that it be connected to the appropriate 'P-OCC-ENTRY' and 'P-OCC-CENS-ENTRY' records in the 'P-OCC-SET' and 'P-OCC-CENS-SET', respectively. If this is the first member record in either of the two sets then it will also be necessary to create the appropriate owner record.

I have now completed my outline examination of the steps which are involved in the simplest possible linkage operation for census household data: the person-based linkage of an occupant who is recorded as being not related to the head of the household. I shall now progressively move on to the much more complicated family-based linkage operations, i.e. those which involve the inter-related members of the household. Each such operation involves two distinct stages. The first is concerned with the attempt to locate the person in the
Population Database. The second is concerned with the way in which the required inter-person links are to be handled.

Firstly, let us consider the linkage of the head of the household. If the primary 'P-FAMILY' record for the household has already been located then access to the head's 'P-PERSON' record is readily obtained by locating the owning record in the 'P-FATHER-SET'. If the primary 'P-FAMILY' record has not been located then the subsequent strategy differs in only minor ways from the linkage of a 'non-related' occupant. If the head cannot be located then a 'P-PERSON' record must be created for him or her. The only significant additional operation is that it will be necessary to note the identity of the head's 'P-PERSON' record (14). It is this record and also the primary 'P-FAMILY' record which will provide the main genealogical reference points when the occupant record for each relative is being linked.

Let us next consider the linkage of the wife of the head of the household. Once again, as for the head, if the primary 'P-FAMILY' record has already been located then the wife's 'P-PERSON' record can readily be obtained, in this case by locating the owning record in the 'P-MOTHER-SET'. If the primary 'P-FAMILY' record has not been located then the subsequent procedure is similar to that already outlined for a married, unrelated occupant. And, in the same way, if the wife cannot be located then a 'P-PERSON' record must be created for her. If there is no primary 'P-FAMILY' record for the present husband and wife, i.e. they are not currently linked in the database, then an
appropriate link should now be created. However, this operation will not necessarily require the creation of a new 'P-FAMILY' record. For example, if another census household record (either earlier or later) for this family has previously been linked, but one where the father, say, was absent, then there will exist a 'fatherless' 'P-FAMILY' record, serving perhaps to connect the mother to the children. For the present linkage operation, where the father is present, it is necessary merely to locate the appropriate 'P-FAMILY' record and connect the father to it in the 'P-FATHER-SET'. The method of locating candidate 'P-FAMILY' records is to search for any member records both in the head's 'P-FATHER-SET' and in the wife's 'P-MOTHER-SET'.

Where no suitable 'fatherless' or 'motherless' 'P-FAMILY' record can be located it will be necessary to create a new 'P-FAMILY' record in order to provide the required husband-wife link. The task of creating the 'P-FAMILY' record will involve the transference of basic, identifying information from the 'P-PERSON' records for the husband and wife to the 'P-FAMILY' record and then the connection of this record to the Cohort and Name Indexes. The detailed operations are as follows:

- initialise various fields and indicators in 'P-FAMILY'.

- obtain a unique family identifier and store it in 'P-FAMILY'.
store also the name information for the husband and wife, together with their birthplace zone codes.

- transfer the 'P-FAMILY' record to the Population Database, i.e. 'create' it.

- since the year of marriage registration will not be known at this stage it will be necessary to connect the 'P-FAMILY' record to the special-purpose 'P-FAMILY-COHORT' record used for such families (see Section 9.2.2).

- use the husband and wife name codes to locate the corresponding 'P-FAMILY-NAME' record, and connect the 'P-FAMILY' record in the 'P-COUPLE-NAM-SET'.

- use the wife's married name codes to locate the corresponding 'P-PERSON-NAME' record, and connect the 'P-FAMILY' record in the 'P-WIFE-NAME-SET'.

- increment by 1 the 'P-ITEM-CNT' record which is used to count the number of families in the Population Database.

Where it has been necessary to locate or create a 'P-FAMILY' record the final operation will be to connect it to the head's 'P-PERSON' record in the 'P-FATHER-SET' and to the wife's 'P-PERSON' record in the 'P-MOTHER-SET', where such connections are not already present. The 'P-FAMILY' record may now be designated and subsequently
Let us next consider the linkage of a son of the head of the household. The attempt to locate the son in the Population Database is carried out in a similar way to that already outlined for a male, unrelated occupant. However, in this search use may be made of the identity of the primary 'P-FAMILY', where this is known for the current household. Thus, for each candidate 'P-PERSON' record located a check may be made on whether its owning 'P-FAMILY' record in the 'P-CHILD-SET' is also present. If it is and the record is not the current primary 'P-FAMILY' then 'no match' is reported. In other words, one can initially confirm that there is no conflict in the 'family of origin' information for the son and each candidate 'P-PERSON' (15). If the son cannot be located in the Population Database then a 'P-PERSON' record must be created for him, as described above.

Having located or created the required 'P-PERSON' record it is now necessary to make the appropriate parent-child connections, where these are not already present. For a household where there is a primary 'P-FAMILY' it will be necessary merely to connect the 'P-PERSON' record for the child to the 'P-FAMILY' record in the 'P-CHILD-SET'. For a household where there is no primary 'P-FAMILY', i.e. because one of the parents is absent, it will be necessary to take appropriate steps to create a link between the head and son. However, as for the head-wife linkage this will not necessarily require the creation of a new 'P-FAMILY' record. A 'fatherless' or,
indeed, 'parentless' owning 'P-FAMILY' record may already exist for the son, and, if so, it will be appropriate to connect this to the head's 'P-PERSON' record in the 'P-FATHER-SET' (or in the 'P-MOTHER-SET', if the head is female). Equally the head might own a 'childless' 'P-FAMILY' record in the 'P-FATHER-SET', and the appropriate action might be to connect the son's 'P-PERSON' record to it in the 'P-CHILD-SET' (16). Finally, and perhaps more rarely, there could be located both 'P-FAMILY' records: one a 'fatherless' 'P-FAMILY' for the son, and the other a 'childless' 'P-FAMILY' record for the father. In this case the appropriate action might be to merge the two 'P-FAMILY' records (17).

Where it has been necessary to locate or create the 'P-FAMILY' record this record may now be designated and subsequently used as the primary family unit in the household. Thus, for example, it can be used when other children of the head present in the household are subsequently linked.

The linkage of an unmarried daughter is precisely the same as for a son. For a married or widowed daughter for whom it is necessary to create a 'P-PERSON' record there will be an additional step. A special 'P-FAMILY' record will need to be created to connect the 'P-PERSON' record to the appropriate 'P-PERSON-NAME' in the Name Index, i.e. the connection must be made via the 'P-WIFE-NAME-SET'. Where the head is male one can deduce the daughter's maiden name, and so one will also be able to connect her to the Name Index via the 'P-BIRTH-NAME-SET', as for an unmarried daughter.
Let us finally consider the linkage of a brother of the head of the household. Since for this relationship one is primarily interested in the head's family of origin, an attempt is therefore made initially to locate this family, i.e. the owning 'P-FAMILY' record in the head's 'P-CHILD-SET'. I shall refer to this family as the 'secondary' family unit in the household. If the family is present in the Population Database use can be made of it when attempting to locate the brother: the method precisely parallels that used when searching for the son of the head, and when given the identity of the primary 'P-FAMILY' unit (18). Again, where the brother cannot be located a 'P-PERSON' record must be created for him.

Having located or created the required 'P-PERSON' record it is then necessary to make the appropriate sibling connections, where these are not already present. For a household where there is a secondary 'P-FAMILY' it will be necessary merely to connect the 'P-PERSON' record for the brother to the 'P-FAMILY' record in the 'P-CHILD-SET'. For a household where there is no secondary 'P-FAMILY' it will be necessary to take appropriate steps to create a link between head and brother. Again, as for the head-wife linkage this will not necessarily require the creation of a new 'P-FAMILY' record. The brother may already have an owning 'P-FAMILY' record in the 'P-CHILD-SET': if so, the appropriate action would be to link the head's 'P-PERSON' record to it in the 'P-CHILD-SET'. Where, instead, the 'P-FAMILY' record needs to be created the 'P-PERSON' records for both head and brother will need to be connected to it in the 'P-CHILD-SET'.

The 'P-FAMILY' record, whether located or newly created, may now be designated and subsequently used as the secondary family unit in the household. Thus, for example, it can be used when other siblings of the head present in the household are subsequently linked.

The logic required for linking a sister can have additional complexity, as was the case for the linkage of a daughter. For an unmarried sister the logic will be precisely the same as for a brother. But for a married or widowed sister for whom it is necessary to create a 'P-PERSON' record then it will also be necessary to make appropriate connections between the 'P-PERSON' record and the Name Index, via a specially created 'P-FAMILY' record.

The strategy for linking the remaining relationships, viz. parents and parents-in-law, are similar to the strategies already described. Thus, an attempt must first be made to locate the person in the Population Database, using where appropriate and where known, the identities of the primary and secondary family units. Once found, or alternatively created, the required 'P-PERSON' record will then be linked to the 'P-PERSON' record for the head (or his wife in the case of an in-law relationship) by appropriate set connections and an intermediate 'P-FAMILY' record. A newly created 'P-PERSON' record for a mother or mother-in-law will need to be connected to the Name Index via the 'P-WIFE-NAME-SET' and a specially created intermediate 'P-FAMILY' record.
Before proceeding to consider the linkage of parish register records I shall review the family-based strategies which are involved in the linkage of census household data. It should be clear that the process of transposing and linking information from the essentially dominant 'person-within-household' grouping to the 'person-within-family' grouping is complex and requires highly intricate sub-strategies appropriate to the various household familial relationships. Indeed, such are the difficulties, that it has seemed worthwhile to provide facilities only for the most common relationships. Others, such as 'grandfather', 'uncle' and 'cousin' are at present regarded as being much too ambiguous for handling in a consistent and reliable way.

10.1.5 Linking Parish Register Records

The linkage of parish register records presents, in general, quite different problems from those of linking census household records, and in this section I shall outline the nature of these differences. I shall then devote the following three sections to a detailed examination of the strategies for linking the individual parish register record types, viz. marriages, baptisms and burials, respectively.

The most fundamental difference between the linkage of census household and parish register records concerns the requirement for information transformation as the records pass to the Population
Database. Thus, in comparison with census household records, parish register records require very little transformation. For example, each pre-1837 marriage record effectively provides a description of a particular familial relationship, i.e. a link between a husband and a wife. When information from such a record is inserted in the Population Database it effectively retains the same structure. What little transformation does occur consists merely of record disaggregation. Thus, a marriage record will essentially need to be transposed into five records: 'P-PERSON' and 'P-EVENT' records for the husband and wife, together with an inter-connecting 'P-FAMILY' record. In this important respect, therefore, the linkage of parish register records presents a strategically simpler task.

The other major difference between the linkage of census household and parish register records concerns the information content of the records. For each individual in a post-1841 census household record there is usually included, in addition to the name, the marital status, age and birthplace. Such information will normally enable inter-census linkages to be carried out with minimal ambiguity, except in more populous communities. By contrast, individuals identified in parish register records rarely have such a range of identifying information included. Thus, if we analyse the records used in the present research we discover the following:

- marital status is not included in burial records. It is, however, included in marriage records and, by implication, for the parents in legitimate baptism records.
age is included consistently only in burial records. Post-1837 marriage records sometimes contain an actual age, but also frequently just the indication 'full age' or 'minor'. Baptism records will in most cases contain implicit age information for the person baptised, but the presence of a birth-baptism interval and the occasional occurrence of adult baptism introduces uncertainty.

- birthplace information is not included explicitly in any of the records. Residence information in a baptism record may, however, provide approximate birthplace identification for the person baptised, although there will be a similar uncertainty here as there is for the age information.

Two conclusions emerge from this analysis. In the first place, the method of identification in parish register records is inferior to that in census household records. And in the second place, there is no consistent method of identification. For a parent identified in a baptism record, for example, there is marital status (implicitly) but not age specified; for a person identified in a burial record, in complete contrast, there is age but not marital status specified.

These deficiencies must inevitably have a fundamental impact on the design of the corresponding linkage strategies. Clearly, the logic must be capable of handling a multiplicity of information-matching situations: this contrasts sharply with the linkage of census data, where the uniformity of structure for each of
the censuses facilitates a simple comparison of parallel information fields in two records. In view of the relative inferiority of the method of identification in parish register records there will clearly be a requirement to make as full use as possible of the information which is available, both in the record to be linked and in the information which is already present in the Population Database.

The strategies which have been designed with these considerations in mind are described in the next three sections. Overall, the strategies are complex and intricate, and proposals for developing a more unified strategy are discussed at the end of the chapter. However, it should be recognised that the major reason for the present complexity is the complexity which is inherent in the problem itself. Thus, for example, in the process of linking a record for a 'JOHN SMITH' it may be necessary to match the details with a 'JOHN SMITH' who is already present in the Population Database. Unfortunately, the information describing the 'JOHN SMITH' in the database can exist in an enormous number of states, depending on which records have been previously linked. There may, for example, be information derived only from a single census event, and with no connections to other relatives. At the other extreme there may be information derived from several census and other events, and with connections to parents, spouse(s) and children. The linkage strategy must clearly be designed to handle both of these linkage situations, as well as all the possible intermediate ones. It is not surprising, therefore, that in these circumstances the logic which results is complex and intricate.
Finally, it should be added that the matching strategies which are described have functioned successfully with the test data used, but that it is recognised that some refinement would probably be necessary were the system to be used with a larger dataset.

10.1.6 Linking Marriage Records

I shall now explore the design of a suitable strategy for linking marriage records, and shall begin by considering the linkage of a pre-1837 marriage record. Such a record will be held in the Source Database as an 'S-MARRIAGE' record (see Figure 8.6), and it will contain for each marriage partner the following information in internal coded form:

- forename, surname and initials
- marital status
- residence
- signature/'X' indicator (i.e. specifying whether the person was or was not able to sign their name)

The record will also contain the unique source record identifier, the date and place of the marriage and an indication of whether the marriage ceremony was conducted by banns or by licence.

It is a relatively simple matter to define the complete set of transformations which must occur when such a record is linked. Thus, the initial task will be to locate or create in the Population Database 'P-PERSON' records for the marriage partners and an
associated 'P-FAMILY' record to hold information about their marriage union. This 'P-FAMILY' record will need to be connected to the groom's 'P-PERSON' record in the 'P-FATHER-SET' and to the bride's 'P-PERSON' record in the 'P-MOTHER-SET'. Finally, it will be necessary to create a 'P-EVENT' record for each of the marriage partners and to connect it in their 'P-LIFE-EVENT-SET'.

As for the linkage of a census household the most appropriate first step when linking a marriage is to attempt to locate the corresponding family unit in the Population Database. The strategy therefore proceeds in a very similar fashion. Thus, the husband and wife name codes from the 'S-MARRIAGE' record are used in combination to attempt to locate the corresponding 'P-FAMILY-NAME' record in the Name Index. If it does not exist then their 'P-FAMILY' record also does not exist, and so the search can be terminated. If it does exist then the 'P-FAMILY' records which are members of the corresponding 'P-COUPLE-NAM-SET' will be interrogated in turn until one is located whose details satisfy the following criteria:

1. it has no marriage record already associated with it.

2. the date of birth of the first child (where specified in the 'P-FAMILY' record) is not more than two years prior to the date of the marriage (19).

3. the dates of birth of each marriage partner (where specified in the 'P-FAMILY' record) do not imply an age at marriage of
Where the attempt to locate the 'P-FAMILY' record is unsuccessful this could be because the required family exists but is not presently connected to both parents, i.e. it is a 'fatherless' or 'motherless' 'P-FAMILY'. The only way to find such a family is to search the Population Database via the groom's name in isolation and then via the bride's married name in isolation. Thus, one would locate the 'P-PERSON-NAME' for the groom and obtain each 'P-PERSON' in the 'P-BIRTH-NAME-SET'. For each 'P-PERSON' one would then interrogate the 'P-FAMILY' records in the 'P-FATHER-SET'. The criteria for a satisfactory match would be as described above, and in addition that the 'P-FAMILY' record was not linked in a 'P-MOTHER-SET'. When searching via the bride's married name one would go directly from the appropriate 'P-PERSON-NAME' record via the 'P-WIFE-NAME-SET' to the candidate 'P-FAMILY' records. The criteria for a satisfactory match would be again as described above, except that in this case the 'P-FAMILY' record must not be linked in a 'P-FATHER-SET'.

If a suitable 'P-FAMILY' record can still not be located then it will be necessary to create one, and the steps involved will closely parallel those for a census family, as described above. The 'P-FAMILY' record, whether located or newly created, can then be used as the primary family unit in the subsequent linkage operations.

The next step in the operation is to process the details concerning the groom. The essential aim is to locate or create a
corresponding 'P-PERSON' record for the groom and attach a 'P-EVENT' record for the marriage in his 'P-LIFE-EVENT-SET'. If the primary family unit has been successfully located and it is not 'fatherless' then the 'P-PERSON' record for the groom can be obtained by locating the owner of the primary 'P-FAMILY' in the 'P-FATHER-SET'. Where this is not possible an attempt must be made to search for the groom in the Population Database. The Name Index is therefore used in the normal way to locate the candidate 'P-PERSON' records for the groom. A matching record will be required to satisfy the following criteria:

1. the computed date of birth (where present in the 'P-PERSON' record) must not imply an age at marriage of less than 12 or greater than 77.

2. the computed date of death (where present in the 'P-PERSON' record) must not be before the date of the marriage.

The remaining criteria must be satisfied by each event record connected to the 'P-PERSON' record in the 'P-LIFE-EVENT-SET'.

3. if the groom is a bachelor then for each event which occurred before the marriage date the marital status must not be 'married' or 'widowed'.

4. for each event which occurred after the marriage date the marital status must not be 'single'.
5. if the groom is a bachelor then there must be no previous marriage event.

If the attempt to locate the groom is unsuccessful then it will be necessary to create a 'P-PERSON' record for him. The steps involved closely parallel those already described for a census occupant, the main difference being that in this case no date of birth or birthplace information will be available and so the 'P-PERSON' record will need to be linked to the special-purpose 'P-PERSON-COHORT' record in the Cohort Index, i.e. the groom cannot be placed in a genuine cohort.

The next step is to connect the primary 'P-FAMILY' record to the groom's 'P-PERSON' record in the 'P-FATHER-SET', where the connection is not already present. Finally, it is necessary to create a marriage 'P-EVENT' record for the groom and connect it in his 'P-LIFE-EVENT-SET'.

A similar sequence of operations must now be carried out to process the details concerning the bride. The only major strategical difference concerns the connection of the bride's 'P-PERSON' record to the Name Index. A pre-1837 marriage record normally contains the bride's pre-marriage name explicitly, and it will also implicitly contain the bride's married name. This therefore enables the bride's 'P-PERSON' record to be connected to the Name Index via two routes (20). The connection of the primary 'P-FAMILY' to the appropriate 'P-PERSON-NAME' record in the 'P-WIFE-NAME-SET' gives a
route via the married name. Where the bride is a spinster it is possible to include a second route via the 'P-BIRTH-NAME-SET' for the maiden name. Alternatively, where the bride is a widow it is possible to include the second route via an additional 'P-FAMILY' record for the former married name. In all cases the necessary records are created and set connections made only where they are not already present.

The final operation when linking a pre-1837 marriage record is to copy details from the record to the primary 'P-FAMILY' record. Such details will include the marriage record unique identifier and the date and place of the marriage, both in internal coded form.

The linkage of post-1837 marriage records employs the same overall strategy as outlined above, except that in this case it is possible to make use of the additional age information which is normally present (21). The age information is first used when the attempt is being made to locate a suitable primary 'P-FAMILY' for the marriage partners. For each candidate 'P-FAMILY' record the parent 'P-PERSON' records in the 'P-FATHER-SET' and 'P-MOTHER-SET' are located, where present. And then for each parent for whom the computed date of birth is specified a check is made on whether it is in agreement with the corresponding age information from the marriage record. Agreement is accepted in the following circumstances:

1. if a true age is given in the marriage record then it must be within 4 years of the age as calculated from the computed date
of birth in the 'P-PERSON' record.

2. if the age in the marriage record is specified as 'full age' or '21' then the age as calculated from the computed date of birth must be greater than 18. (22)

3. if the age in the marriage record is specified as 'minor' then the age as calculated from the computed date of birth must be less than 23. (23)

The above age criteria are also used in a similar way later in the linkage strategy in the situation where it is necessary to attempt to locate a 'P-PERSON' record for the groom and/or bride. This situation arises where no primary 'P-FAMILY' can be located or where one is located, but with a missing parent.

Finally, the presence of occupation information for the marriage partners in a post-1837 marriage record requires a minor, additional enhancement to the linkage strategy. As each 'P-EVENT' record is created and connected in its 'P-LIFE-EVENT-SET', it will also be possible to connect it to the Occupation Index: this will be achieved by making a corresponding connection in the appropriate 'P-OCC-SET'. 
10.1.7 Linking Baptism Records

Baptism records, in general, have a structural similarity with pre-1837 marriage records. Thus, they normally contain the names of a husband and wife, and they also lack associated age information. They differ primarily in three respects. Firstly, and quite obviously, a baptism record contains the details of a child of the marriage. Secondly, a baptism record normally does not contain the pre-marriage name (usually maiden name) of the wife; a marriage record almost always does have this name. Finally, a baptism record can relate to a legitimate or an illegitimate birth event; a marriage record relates exclusively to a legitimate event.

Let us initially consider the linkage of a baptism record for a legitimate birth, and one in which the maiden name of the wife is not included. Such a record will be held in the Source Database as an 'S-BIRTH' record (see Figure 8.6), and it will normally contain the following information in internal coded form:

- the unique source record identifier
- the forename, initials and gender of the person baptised
- date and place of baptism
- computed date of birth of the person baptised (the computation being based on the user-specified birth-baptism interval)
  - or an indication of an adult baptism
- residence
- the father's forename, surname, initials and occupation
- the mother's forename and initials
As for the linkage of a marriage record, it is a relatively simple matter to define the complete set of transformations which must occur when such a record is linked. Thus, the initial task is to locate or create in the Population Database 'P-PERSON' records for the parents and child and an associated 'P-FAMILY' record to hold information about the marriage union. This 'P-FAMILY' record will need to be connected to the father's 'P-PERSON' record in the 'P-FATHER-SET', to the mother's 'P-PERSON' record in the 'P-MOTHER-SET' and to the child's 'P-PERSON' record in the 'P-CHILD-SET'. In addition it will be necessary to create a 'P-EVENT' record for each person and to connect it in their 'P-LIFE-EVENT-SET'.

Once again, as for the linkage of a census household or marriage record, the most appropriate first step when linking a baptism is to attempt to locate the corresponding family unit in the Population Database. The strategy therefore proceeds in a very similar fashion. Thus, the father and mother name codes from the 'S-BIRTH' record are used in combination to attempt to locate the corresponding 'P-FAMILY-NAME' record in the Name Index. If it does not exist then the required 'P-FAMILY' record also does not exist, and so the search can be terminated. If it does exist then the 'P-FAMILY' records which are members of the corresponding 'P-COUPLE-NAM-SET' will be interrogated in turn until one is located whose details satisfy the following criteria:

1. if the 'P-FAMILY' record contains a marriage date then this must not be later than the computed date of birth of the
child.

2. the computed date of birth of the child must not be more than 35 years after the computed date of birth of the first child (where specified in the 'P-FAMILY' record) and it must not be more than 35 years before the computed date of birth of the last child (where specified in the 'P-FAMILY' record).

If an acceptable 'P-FAMILY' record, i.e. the primary family, is located then it can be used as the 'reference' family for the subsequent linkage operations for the current baptism.

The next step in the strategy is to attempt to locate the child in the Population Database. The Name Index is used in the normal way to locate the candidate 'P-PERSON' records for the child. A matching record must satisfy each of the following criteria:

1. if the primary 'P-FAMILY' is known and this record has either or both of its 'P-PERSON' parent records linked then neither of these records may be accepted as a suitable candidate 'P-PERSON' record (24).

2. the candidate 'P-PERSON' must not already have a baptism associated with it.

3. if the 'P-PERSON' record has an owning 'P-FAMILY' in the 'P-CHILD-SET' then there must be no mismatch between the major
forename codes for the father and mother contained in this record and the corresponding codes in the 'S-BIRTH' record.

4. if the 'P-PERSON' record has already had a burial associated with it then the date of burial must not be prior to the date of baptism.

5. the 'P-PERSON' record must not indicate an illegitimate birth.

6. if the baptism is not specified as being an adult one then any birth information contained in the 'P-PERSON' record must be compatible with the corresponding information in the baptism record. There is assumed to be compatibility if the date of birth computed from the baptism date and the date of birth in the 'P-PERSON' record agree to within two years. Compatibility is also assumed if the dates agree to within 2-6 years and there is no disagreement in the birthplace zone codes.

7. if the candidate 'P-PERSON' record owns one or more 'P-FAMILY' records in the 'P-FATHER-SET' or 'P-MOTHER-SET' then a check is made that the date information in each 'P-FAMILY' record does not conflict with the computed date of birth of the child. If the marriage date is specified then this must not imply an age at marriage for the child of less than 15. Similarly if the date of birth of the first child is specified then this must be at least 15 years after the date of birth of
the child in the baptism. And if the date of birth of the last child is specified then this must be no more than 77 years after the birth of the child in the baptism if male, and no more than 52 years if female.

If the attempt to locate the child's 'P-PERSON' record is successful and also the primary 'P-FAMILY' is known then a further check must now be carried out. Where the 'P-PERSON' record has an owning 'P-FAMILY' record in the 'P-CHILD-SET' it will sometimes be found that this record is not also the primary 'P-FAMILY' (25). If this is the case then it will be necessary for the information from the two 'P-FAMILY' records to be merged into one and for the other record to be destroyed.

If the attempt to locate the child is unsuccessful then it will be necessary to create a new 'P-PERSON' record. The steps involved closely parallel those already described for a census occupant. If the 'S-BIRTH' record does not indicate an adult baptism then a computed date of birth is calculated from the baptism date and the birthplace is taken to be the residence given in the baptism record. Provided with this estimated birth information one will then be able to link the 'P-PERSON' record to the appropriate 'P-PERSON-COHORT' record in the Cohort Index. For an adult baptism no birthplace information, and often no date of birth information, is available and so the 'P-PERSON' record will need to be linked to the special-purpose 'P-PERSON-COHORT' record in the Cohort Index, i.e. the person cannot be placed in a genuine cohort.
The next step is to deal with the child-family connection. If the primary 'P-FAMILY' record has been located then it will be possible to connect the child's 'P-PERSON' record to it in the 'P-CHILD-SET', where the connection is not already present. Conversely, if the primary 'P-FAMILY' record has not been located, but the 'P-PERSON' record has an owning 'P-FAMILY' in the 'P-CHILD-SET' then this 'P-FAMILY' can be located and designated as the primary family. The final operation in the linking of the child is to create a baptism 'P-EVENT' record and connect it in the child's 'P-LIFE-EVENT-SET'.

The remaining operations in the linkage of a legitimate baptism are concerned with the processing of the details for the father and the mother. Let us first consider the details for the father. If the primary family unit has been successfully located and it is not 'fatherless' then the 'P-PERSON' record for the father can be obtained by locating the owner of the primary 'P-FAMILY' in the 'P-FATHER-SET'. Where this is not possible an attempt must be made to locate the father in the Population Database. The Name Index is used in the normal way to locate the candidate 'P-PERSON' records for the father. A matching record must satisfy the following criteria:

1. where computed dates of birth are available for the candidate father (i.e. from the 'P-PERSON' record) and for the child baptised then the difference between the dates must be such that the father is not less than 15 years and not more than 77 years older than the child.
2. if the 'P-PERSON' record has a burial record associated with it then the computed date of death for the candidate father must not be more than three years prior to the date of the baptism (26).

The final criterion must be satisfied by each event record connected to the 'P-PERSON' record in the 'P-LIFE-EVENT-SET'.

3. if the event date is after the date of the baptism then the marital status must not be 'single'.

If the attempt to locate the father is unsuccessful then it will be necessary to create a 'P-PERSON' record for him. The steps involved closely parallel those for a census occupant: but because of the absence of date of birth and birthplace information the 'P-PERSON' record will need to be linked to the special-purpose 'P-PERSON-COHORT' record in the Cohort Index, i.e. the father cannot be placed in a genuine cohort.

The next step is to make the appropriate familial connections. If the primary 'P-FAMILY' record has been located then it is necessary merely to connect it to the father's 'P-PERSON' record in the 'P-FATHER-SET', where the connection is not already present. If the primary 'P-FAMILY' record has not been located then it is necessary to take appropriate steps to create a link between father and child. The strategy for doing this precisely parallels that for linking a child to the head of a census household, as described in Section 10.1.4.
Finally, it is necessary to create a 'P-EVENT' record for the father and connect it in his 'P-LIFE-EVENT-SET'. This 'P-EVENT' will register the fact that on the specified date the person was identified as a father in a baptism record. Where the father's occupation has been specified it will also be possible to connect the 'P-EVENT' record, in the usual way, to the Occupation Index.

A similar sequence of operations must now be carried out to process the details for the mother. The first significant alteration in strategy occurs where it is necessary to attempt to locate the mother in the Population Database. Since only her married name is known the search for candidate 'P-PERSON' records must be made via the 'P-WIFE-NAME-SET' and the intermediate 'P-FAMILY' records. A matching 'P-PERSON' record must satisfy the following criteria:

1. where computed dates of birth are available for the candidate mother (i.e. from the 'P-PERSON' record) and for the child baptised then the difference between the dates must be such that the mother is not less than 15 years and not more than 52 years older than the child.

2. if the 'P-PERSON' record has a burial record associated with it then the computed date of death for the candidate mother must not be more than three years prior to the date of the baptism (27).
3. the intermediate 'P-FAMILY' record must not indicate an illegitimate union.

4. where the intermediate 'P-FAMILY' record has a code for the husband's major forename this must be identical to the major forename code for the father in the baptism record.

The final criterion must be satisfied by each event record connected to the 'P-PERSON' record in the 'P-LIFE-EVENT-SET'.

5. if the event date is after the date of the baptism then the marital status must not be 'single'.

If the attempt to locate the mother is unsuccessful then it will be necessary to create a 'P-PERSON' record for her. The remaining steps in the linkage are similar to those for the father. Thus, the mother's 'P-PERSON' record is connected to the primary 'P-FAMILY' in the 'P-MOTHER-SET', where the connection is not already made. Finally, a 'P-EVENT' record is connected in her 'P-LIFE-EVENT-SET'. However, since occupation information is not normally provided for the mother it will not also be possible to connect this 'P-EVENT' record to the Occupation Index, as was the case for the father.

The linkage of an illegitimate baptism record is a rather simpler task to perform than the linkage of a legitimate one since there are only two people involved, the mother and the child. When such a record is linked it is necessary to locate or create in the Population
Database 'P-PERSON' records for the mother and child and an associated 'P-FAMILY' record to hold information about the illegitimate union. This 'P-FAMILY' record will be connected to the mother's 'P-PERSON' record in the 'P-MOTHER-SET' and to the child's 'P-PERSON' record in the 'P-CHILD-SET': there will be no connection in the 'P-FATHER-SET'. In addition, it will be necessary to create a 'P-EVENT' record for each person and to connect it in their 'P-LIFE-EVENT-SET'.

The first step is to attempt to locate the child in the Population Database. The strategy is precisely the same as for the legitimate baptism, except that the fifth criterion (viz. that the candidate 'P-PERSON' record must not indicate an illegitimate birth) must clearly be ignored. If the attempt to locate the child is unsuccessful then it will be necessary to create a new 'P-PERSON' record, as before. The 'P-PERSON' record, whether located or newly created, will be marked as 'illegitimate' by setting the appropriate indicator value. The final operation in processing the information for the child is to create a baptism 'P-EVENT' record and connect it in the child's 'P-LIFE-EVENT-SET'.

The next step is to attempt to locate the mother in the Population Database. With the present linkage strategy the assumption is made that the mother is unmarried (28). The search for candidate 'P-PERSON' records is therefore made, as for the child, via the appropriate 'P-BIRTH-NAME-SET'. A matching 'P-PERSON' record must satisfy the following criteria:
1. Where computed dates of birth are available for the candidate mother (i.e. from the 'P-PERSON' record) and for the child baptised then the difference between the dates must be such that the mother is not less than 15 years and not more than 52 years older than the child.

2. If the 'P-PERSON' record has a burial record associated with it then the computed date of death for the candidate mother must not be more than three years prior to the date of baptism.

   The final criterion must be satisfied by each event record connected to the 'P-PERSON' record in the 'P-LIFE-EVENT-SET'.

3. If the event date is before the date of the baptism then the marital status must not be 'married'.

If the attempt to locate the mother is unsuccessful then it will be necessary to create a 'P-PERSON' record for her. The final operation in processing the information for the mother is to create a 'P-EVENT' record and to connect it in her 'P-LIFE-EVENT-SET'. In many instances occupation information will be provided for the mother: this is frequently of the form 'SINGLE WOMAN' or 'SPINSTER'. In such cases it will therefore be additionally possible to connect the 'P-EVENT' record to the Occupation Index, in the usual way.
The final linkage operation is to connect the 'P-PERSON' records for the mother and child via an intermediate 'P-FAMILY' record, where the connection is not already made. Once again, the strategy for doing this will precisely parallel that for linking a child to the head of a census household, as described in Section 10.1.4. But it should be noted that in this case the 'P-FAMILY' record must be marked as 'illegitimate' by setting the appropriate indicator value.

10.1.8 Linking Burial Records

Post-1812 burial records are the simplest records which the linkage system is required to handle. They contain information about only one person, the deceased, and apart from the person's name and the date they contain only two other items of information: age and residence. They are moderately simple to link, involving only person-based linkage operations, although the absence of marital status information can cause problems when linking records for women.

Let us initially consider the linkage of a burial record for a male person. Such a record will be held in the Source Database as an 'S-DEATH' record (see Figure 8.6), and it will normally contain the following information in internal coded form:

- the unique source record identifier
- the forename, surname, initials and gender of the person buried
- date and place of burial
computed date of birth of the person buried
computed date of death of the person buried
residence

The set of transformations which must occur when such a record is linked are very simple to define. Thus, the initial task is to locate or create in the Population Database a 'P-PERSON' record for the person deceased. The only other task is to create a burial 'P-EVENT' record for them and to connect it in their 'P-LIFE-EVENT-SET'.

Let us first consider how the person is searched for in the Population Database. The Name Index is used in the normal way to locate the candidate 'P-PERSON' records. A matching record must satisfy each of the following criteria:

1. it must not already have a burial associated with it.

2. the computed date of birth (where present in the 'P-PERSON' record) must not be after the date of burial and it must be within four years of the computed date of birth of the deceased.

3. the date of the chronologically last event record in the candidate's 'P-LIFE-EVENT-SET' must not be after the burial date.
Where the family-of-origin for a candidate 'P-PERSON' record is present in the Population Database (i.e. the owning 'P-FAMILY' in the 'P-CHILD-SET') then the following additional criteria must also be satisfied:

4. the date of marriage, if present, must not be after the date of burial.

5. the computed date of birth of the deceased must not be more than 35 years after the computed date of birth of the first child (where specified in the 'P-FAMILY' record) and it must not be more than 35 years before the computed date of birth of the last child (where specified in the 'P-FAMILY' record).

6. if the 'P-FAMILY' record has its owning 'P-PERSON' record in the 'P-FATHER-SET' and this record has a computed date of birth for the father then this date must not predate the computed date of birth of the deceased by less than 15 years or more than 77 years.

7. if the 'P-FAMILY' record has its owning 'P-PERSON' record in the 'P-MOTHER-SET' and this record has a computed date of birth for the mother then this date must not predate the computed date of birth of the deceased by less than 15 years or more than 52 years.
If the candidate 'P-PERSON' record owns one or more 'P-FAMILY' records in the 'P-FATHER-SET' (i.e. corresponding with his family or families of procreation) then it is necessary to check that the date information in each 'P-FAMILY' record does not conflict with the computed date of birth of the deceased. The following criteria must be satisfied:

8. if the marriage date is specified then this must not imply an age at marriage for the deceased of less than 15 years.

9. if the date of birth of the first child is specified then this must be at least 15 years after the date of birth of the deceased.

10. if the date of birth of the last child is specified then this must not be more than 77 years after the date of birth of the deceased.

If the attempt to locate the deceased is unsuccessful then it will be necessary to create a new 'P-PERSON' record. The steps involved closely parallel those for a census occupant: but because of the absence of birthplace information (and even though a computed date of birth is available) the 'P-PERSON' record will need to be linked to the special-purpose 'P-PERSON-COHORT' record in the Cohort Index, i.e. the person cannot be placed in a genuine cohort.
Since a post-1812 burial record contains information about only one person the linkage of such a record will entail no familial connections, unlike each of the other records which have been considered. The only remaining step, therefore, is to create a burial 'P-EVENT' record and connect it in the deceased's 'P-LIFE-EVENT-SET'.

A similar sequence of operations is carried out when linking a burial record for a female person. The most significant difference occurs when the attempt is made to search for the deceased in the Population Database. Since there is no indication of marital status there will be a corresponding uncertainty about whether the name is a maiden or married name. If it is a maiden name then the search for candidate 'P-PERSON' records should be made via the 'P-BIRTH-NAME-SET'; if it is a married name then the search should be made via the 'P-WIFE-NAME-SET' and the intermediate 'P-FAMILY' records. The appropriate strategy is therefore as follows:

1. If the age of the deceased is no more than 12 years then the name is assumed to be a maiden name and the search is made only via the 'P-BIRTH-NAME-SET'.

2. If the age of the deceased is greater than 12 years but no more than 21 years then it is assumed that the name is probably a maiden name, but that it could be a married name. A search is therefore made first via the 'P-BIRTH-NAME-SET', and if this is unsuccessful then secondly via the 'P-WIFE-NAME-SET'.
3. If the age of the deceased is greater than 21 then it is assumed that the name is probably a married name, but that it could be a maiden name. In this case therefore a search is made first via the 'P-WIFE-NAME-SET', and if this is unsuccessful then secondly via the 'P-BIRTH-NAME-SET'.

The other significant change in the linkage strategy occurs where it is necessary to create a new female 'P-PERSON' record in the Population Database. Since the marital status is uncertain there is a consequent problem about how the 'P-PERSON' record should be connected to the Name Index, i.e. whether via the 'P-BIRTH-NAME-SET' or via the 'P-WIFE-NAME-SET'. In the current implementation the problem is resolved by always connecting via the 'P-BIRTH-NAME-SET', i.e. by assuming that the deceased is unmarried. For a person who is over 12 a special marker is then set in the 'P-PERSON' record to register the fact that there is a potential name ambiguity. However, since with the present linkage strategy burial records are by default the last ones to be linked, there is normally no opportunity for such name ambiguities to be resolved.

10.1.9 The Control Structure for Record Linkage

I shall now examine the design of those parts of the system which actually carry out the record linkage operations. In Section 9.1.2 I described the internal design and control structure of the part of the Record Linkage Subsystem which is responsible for overall control and
part 2 - record linkage strategies

for record selection and display: the relevant modules were illustrated in Figure 9.4. It is now necessary to look at the additional modules in the subsystem which are needed to carry out the linkage functions and organise information in the Population Database. The relevant modules and control structure are shown in Figure 10.1. For simplicity only the major module interconnections are shown.

What these modules are required to accomplish is the total organisation of the information in the Population Database. In the first place they have the responsibility for carrying out the transformations on the individual source records which are selected from the Source Database. For each such record it will be necessary to transpose it into one or more 'person', 'family' and 'event' records, and install and link these records in appropriate relationship to the other records in the Population Database. Particular modules will need to be responsible for organising the Name, Cohort and other indexes and for connecting each 'person' and 'family' record to them. Finally, other modules will be required to supervise the assembly of genealogical structures from the individual 'person' and 'family' records.

The preponderence of database-handling modules in Figure 10.1 confirms my earlier observation that the Record Linkage Subsystem has considerable interaction with the database. Thus, of the 21 modules shown in Figure 10.1 only the two in the bottom right-hand corner, viz. 'T2TIMEMACHINE' and 'T4DATESCOMPARE', do not interface directly with IDMS. Broadly speaking, the modules in the upper half of the
Figure 10.1  The Record Linkage Subsystem - record linkage: structure and flow of control
diagram are concerned with the Source Database and with the handling of the various kinds of source records; the modules in the lower half of the diagram are concerned with the Population Database and with the handling of 'person', 'family' and 'event' records, and the associated indexes.

I shall now examine in detail the way in which the individual modules contribute to the total linkage operation, and shall begin by considering the linkage of census household records for a particular surname. As was described in Section 9.1.2 module 'PZDBPROCESS', shown at the top of Figure 10.1, has overall control for the record display and linkage operations. Having located the appropriate 'N-SURNUM' record in the Source Database (see Figure 8.6) and having established both that there are census records connected to it and that the user has requested that such records are to be linked, module 'PZDBPROCESS' calls module 'SVLINKCENSUS'. Where, for example, the user has requested that only 1851 census records are to be linked 'SVLINKCENSUS' must traverse the two sets 'S-CENS-NAME-SET' and 'S-CENS-LINK-SET' in search of the records for the relevant census year.

Before each selected household is linked it is first necessary to establish whether or not the household has already been linked during an earlier linkage operation. Module 'SSCHECKREC' is invoked to carry out the necessary check. To do this it attempts to create a 'P-SOURCE-PRESENT' record using the unique source record identifier for the household. If the record is found to be already present then
this is reported back to 'SVLINKCENSUS', which then abandons further processing of the current household and moves on to the next one. If the record was not already present then this is reported back, and 'SVLINKCENSUS' proceeds to invoke module 'STCELINK' to supervise the linkage operations.

Module 'STCELINK' is responsible for analysing the 'S-CENS-HOUSEHOLD' record which has been located and for invoking module 'SRSTORECEPOP' to generate a corresponding 'P-CENS-HOUSEHOLD' record in the Population Database. Its other main function is to traverse the 'S-CENS-OCC-SET' so that each 'S-CENS-PERSON' record can be retrieved and the appropriate linkage operations initiated. For each person in the household who is related to the head it provides the appropriate logic for handling the particular family relationship: wife, son, brother, etc. For each individual in the household 'STCELINK' invokes 'SRSTORECEPOP' to create a 'P-CENS-EVENT' record: each such record corresponds with one 'S-CENS-PERSON' record in the Source Database. 'SRSTORECEPOP' invokes module 'JXLINK' when a 'P-CENS-EVENT' record is to be connected to the Occupation Index.

Module 'STCELINK' invokes a number of support modules which have responsibility for handling all accesses to particular groups of records in the database. For example, module 'NWNAMES' controls access to the Name Index. When 'STCELINK' needs to find out whether a particular person is already present in the Population Database it presents to 'NWNAMES' the forename and surname code values, together with other information obtained from the relevant source record.
'NWNAMES' is then responsible for attempting to locate the name in the Name Index and for locating and testing any candidate 'P-PERSON' records. In a similar fashion it is responsible for sifting through candidate 'P-FAMILY' records when presented with name information for a husband and wife. It is additionally responsible for creating records in the Name Index and for connecting 'P-PERSON' and 'P-FAMILY' records to them.

Access to the Cohort Index is handled in a similar way by module 'TZCOHORTS'. It is responsible for creating the necessary records and for connecting 'P-PERSON' and 'P-FAMILY' records to them.

The main organisation of the 'P-PERSON' and 'P-FAMILY' records and their familial interconnections in the 'P-FATHER-SET', 'P-MOTHER-SET' and 'P-CHILD-SET' is handled by two modules: 'PWPOPPROCESS' and 'PVPOPFUNCTION'. 'PWPOPPROCESS' operates at the more primitive level, and is responsible for the initial creation of 'P-PERSON' and 'P-FAMILY' records, and, where necessary, for their destruction (29). Module 'PVPOPFUNCTION' has responsibility for the familial connections between 'P-PERSON' and 'P-FAMILY' records, i.e. for all the processing concerned with insertions in the 'P-FATHER-SET', 'P-MOTHER-SET' and 'P-CHILD-SET'. Not only does it handle the basic operations of connecting parents and children to families, i.e. providing the elementary 'P-PERSON'-'P-FAMILY' links; it also carries out the more complex tasks of connecting two siblings and connecting a parent to a child. In these more complex tasks it must arrange for the location or creation of an intermediate
'P-FAMILY' record to provide the connection between the two 'P-PERSON' records. Module 'PVPOPFUNCTION' may also be invoked to make disconnections between 'P-PERSON' and 'P-FAMILY' records: such operations may be needed when two 'P-PERSON' records are to be merged or when two 'P-FAMILY' records are to be merged.

The handling of all accesses to the 'P-ITEM-CNT' records is carried out by module 'PXCTRPROCESS'. This module arranges for unique six-character identifiers to be obtained for the 'P-PERSON' and 'P-FAMILY' records and it also creates and then updates, as necessary, the five records which contain the counters for persons, families, surnames, person names and family names.

Finally, there are the two additional support modules which make no access to records in the database: 'T2TIMEMACHINE' and 'T4DATESCOMPARE'. Module 'T2TIMEMACHINE', which is also extensively used by the Source Translation Subsystem, is responsible here only for one function: it is required to calculate the year value when presented with an internal coded date. For example, when presented with the computed date code '62731183' it would return the string '1821'. Such a calculation is needed, for example, when a 'P-PERSON' record is to be connected to the appropriate 'P-PERSON-COHORT' record in the Cohort Index.

Module 'T4DATESCOMPARE' is responsible for comparing two dates in internal coded form and for measuring the difference between them. It returns a code value indicating the amount of agreement, and if, for
example, it was presented with the two computed date codes '62731183' and '63249030' it would return a code value indicating that the dates are more than 1 year, but less than 2 years apart. Such comparisons are needed when attempting to find matching 'P-PERSON' records in the Population Database.

The linkage of birth, marriage and death records is carried out in a similar fashion, and corresponding use is made of the various support modules. For example, where marriage records are to be linked module 'USLINKMAR' traverses 'S-MAR-NAME-SET', 'S-MAR-LINK-SET' and 'S-MAR-BR-LNK-SET' to gain access to the required marriage records in the Source Database (see Figure 8.6). As each record is located 'USLINKMAR' invokes module 'SSCHECKREC' to check whether the record has previously been linked. If it has not been linked then 'USLINKMAR' proceeds to invoke module 'UUMARLINK' to supervise the linkage operations. The subsequent actions carried out by 'UUMARLINK' parallel those carried out for census data by 'STCELINK'. For simplicity the connections between 'UUMARLINK' and the various support modules are not included in Figure 10.1.

Module 'PJMERGE' is an additional support module which has responsibility for merging two 'P-PERSON' records and for merging two 'P-FAMILY' records. It may be invoked during the linkage of birth and marriage records.
10.2 AN EXAMINATION OF THE PROBLEM OF NOMINAL AMBIGUITY

Central to my approach to the task of record linkage is the contention that for the kinds of data being used, nominal ambiguity is not a major problem, at least when dealing with smaller communities (30). In order to verify this I have carried out a nominal analysis of the populations of five English communities by examining machine-readable versions of their 1851 census returns (31). My chief objective has been to determine the percentage of the population of each community for whom there was a potential nominal ambiguity problem. In other words, put simply, were there many people in the community called 'John Smith', say, who were approximately the same age?

For each of the communities I used a specially written Fortran program to read the census data file and extract for each individual the following details: surname, forename(s), age, marital status, community of birth, county of birth and household schedule number. These details were written to a file, which was then sorted alphabetically so as to bring together all those sharing the same name. The final task was to search through the file manually so as to locate those individuals who share ambiguous identifying information. People were judged to share ambiguous identifying information if each of the following was found to be true for them:

1. they share the same or similar surnames. For example, 'ARNITT' and 'ARNOT' are treated as synonyms.
2. they share the same or similar forenames. For example, 'MARY' and 'MARIA' are treated as synonyms. But also 'ANN' is treated as a synonym for 'MARY ANN'.

3. they have matching gender.

4. their ages are within 5 years of each other.

The results of the analysis are illustrated in Table 10.1. Consider the first entry in the table, viz. for Bucklebury in Berkshire. Of the total population of 1201, 77 (i.e. 6.4%) are potentially ambiguous according to the above criteria. However, this is a 'worst case' figure, since it does not take into account other identifying information which is also normally present in census household data, particularly that relating to an individual's relatives. I therefore used the household schedule number for each of my potentially ambiguous individuals in order to cross-reference their household information. I then examined the following information (where present) to determine whether ambiguities could be resolved:

- the forenames of one or more parents of the individual
- the forename(s) of the individual's spouse
- the individual's birthplace code (32)
### Table 10.1 Assessment of nominal ambiguity from 1851 census returns

<table>
<thead>
<tr>
<th>Community</th>
<th>Population</th>
<th>Potentially ambiguous persons N</th>
<th>%</th>
<th>Ambiguities resolved by parent name(s)</th>
<th>spouse name</th>
<th>birthplace code</th>
<th>Remaining ambiguous persons N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bucklebury, Berkshire</td>
<td>1201</td>
<td>77</td>
<td>6.4</td>
<td>61</td>
<td>12</td>
<td>0</td>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td>Grinstead, Norfolk</td>
<td>1240</td>
<td>86</td>
<td>6.9</td>
<td>50</td>
<td>27</td>
<td>0</td>
<td>9</td>
<td>0.7</td>
</tr>
<tr>
<td>Dodbrooke, Devon</td>
<td>1300</td>
<td>38</td>
<td>2.9</td>
<td>20</td>
<td>14</td>
<td>2</td>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>Deeping Fen, Lincolnshire</td>
<td>1099</td>
<td>41</td>
<td>3.7</td>
<td>9</td>
<td>13</td>
<td>11</td>
<td>8</td>
<td>0.7</td>
</tr>
<tr>
<td>North Sunderland, Northumberland</td>
<td>1198</td>
<td>108</td>
<td>9.0</td>
<td>62</td>
<td>39</td>
<td>3</td>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td>Mean*</td>
<td>1208</td>
<td>70</td>
<td>5.8</td>
<td>40</td>
<td>21</td>
<td>3</td>
<td>5</td>
<td>0.4</td>
</tr>
</tbody>
</table>

* N.B. Due to rounding, there is a discrepancy between the mean value of N in the penultimate column and the value as derived from the figures in the Mean row.
Table 10.1 shows that of the 77 potentially ambiguous individuals in Bucklebury, 61 become unambiguous when parent forenames are taken into account. Of the remaining 16, 12 become unambiguous when spouse forenames are taken into account. The birthplace codes do not resolve the ambiguity for any of the remainder, and so 4 (i.e. 0.3%) remain ambiguous.

The results are reasonably comparable for the other four communities, although Deeping Fen in Lincolnshire has relatively few individuals (viz. 9) whose ambiguity is resolved by parent forenames and relatively many (viz. 11) whose ambiguity is resolved by birthplace codes. Averaging over all the communities one finds that only 0.4% of the population share ambiguous identifying information. Even this may be slightly pessimistic, since the age range specified, viz. 5 years, is somewhat generous, given that age reporting is generally thought to be reliable and correct to within two years (33). It is also noteworthy from the mean figures that the simple zoning scheme for birthplace matching succeeds in making a small but significant contribution to the resolution of ambiguities. Thus, on average, of 8 remaining ambiguous persons, 3 (i.e. approximately 35%) are resolved by the birthplace code.

If it is assumed that for the remaining ambiguous individuals it is possible to make the correct linkage in 50% of the cases (34), then one can conclude that the number of incorrect linkages which are likely to occur for such a population will be in the region of one or two in every 1000. This is consistent with Levine's findings for
An examination of the 27 individuals who remain ambiguous reveals that 10 of these consist of five pairs of sisters sharing similar names: for three of the pairs the names are 'ELIZA' and 'ELIZABETH'. The remaining 17 individuals are mostly household occupants who are unrelated to the head, e.g. lodgers and servants. As such, they usually have no co-residing relatives, and so they lack the additional identifying information which such people provide.

In this section I have demonstrated that nominal ambiguity is not a significant problem when linking census data for communities of around 1200 people. However, is it likely that the same conclusion will hold for other data, and, in particular, for the kinds of parish register data which were used? There is reasonable confidence that this will indeed be the case and that a comparable level of linkage accuracy can be achieved. In the first place, a significant advantage of the present linkage strategy is that by carrying out the prior linkage of the census records one is able to establish the main family groupings in the population and set up a large body of information against which to match each new record (36). But secondly, each of the parish register records has particular strengths when considering its linkage to the data already held in the Population Database. Thus, each marriage record includes the names of two people (i.e. husband and wife), and each legitimate baptism record normally includes the names of three people (i.e. father, mother and child). By subjecting such records to linkage strategies designed in
accordance with the family-based linkage concepts which I have
developed, there is therefore little scope for linkage ambiguity. The
only records for which one is essentially reduced to using
person-based linkage strategies are the burial records, and that is
because such records lack the names of any relatives. They do,
however, carry the age of the deceased, and so the scope for ambiguity
is low. In addition, since these records are linked last of all, the
personal 'dossiers' which will by now have been established in the
Population Database will constrain severely the linkages which are
permissible, and so false links should be avoided.

10.3 THE RECORD LINKAGE STRATEGY – A CRITICAL RE-APPRAISAL

The record linkage strategy described in this chapter has been
used successfully to link nineteenth century census and parish
register records for the small parish of Elwick Hall in County Durham.
A critical analysis of the results of this linkage is provided in
Chapter 12. I shall now consider to what extent the linkage strategy
would be suitable for larger populations, and, more generally, how
appropriate to the linkage problem are the various design features of
the system.

Firstly then, what would be the repercussions of using the
linkage system for a larger population, possibly as large as 10,000,
the maximum which was originally envisaged? As far as the user of the
system was concerned he would, of course, be required to prepare and submit much larger quantities of source records. But also he would need to increase the sizes of the five name directories so that they would be able to handle the additional name information present in the new records. Fortunately this task would not increase linearly with the population size, since each addition of new records would tend to contain diminishing numbers of name fields which were not already present in the directories.

More significantly, how would the linkage strategies cope with an increased nominal record universe? In Section 10.2 I demonstrated that for the linkage of census data for communities of around 1200 one might expect only one or two linkages in every 1000 to be incorrect. Scaling this up to a population size of 10,000 the expected error rate might still then be only around one or two in every 100 linkages. This result is possibly over-pessimistic, since with larger populations there will tend to be a larger universe of forename-surname combinations, and so one would not expect the problem of nominal ambiguity to increase linearly with population size. However, there may be other, more subtle factors which will tend to counterbalance this effect and cause the nominal ambiguity to increase more than linearly with population size (37).

Further confidence in the capacity of the linkage strategy to handle larger populations is provided by the fact that the various record matching algorithms which were employed were installed and made operational without the need for significant adjustment: there would
appear therefore to be considerable scope for modification and refinement, should this prove necessary. And the ease with which the matching algorithms were made operational would appear to support my view that it is the organisation of family-based linkage and the handling of structurally complex data, rather than nominal ambiguity, which pose the severest problems when linking nineteenth century nominal records.

The crucial requirement of any record linkage strategy is that, regardless of its internal mode of operation, it should produce the 'correct' linkages. According to this measure the record linkage strategy which has been developed performs satisfactorily, this conclusion being amply supported by analyses contained in Chapter 12.

In the following sections I shall review the design of the present strategy and make some proposals about possible, future enhancements.

10.3.1 An Appraisal of the Design of the Present Linkage Strategy

The system as presently implemented provides an effective and easy to use means of achieving the family-based record linkage of nineteenth century demographic data. It is able to process a range of types of data, and the method of handling these is mutually consistent: the interface to the user is therefore kept simple and coherent. The convergence of record types to a small number of basic types (e.g. the convergence of all marriage records to one internal format) is a significant and valuable feature, and it ensures that the
system will be able to incorporate new variants of the basic types with minimal inconvenience.

In addition, the centralised handling of name information (e.g. surnames, Christian names, occupations, etc.) by a series of directories provides a consistent and very successful method of handling the records at the individual field level. Again, this simplifies the incorporation of new types of records, and it should also facilitate the use of the system for handling data from other geographical areas and even from other countries. (By contrast, when using an algorithmic solution (e.g. Soundex) one would need to modify it for each new area encountered.)

The central task of implementing the family-based linkage of household census and parish register data presented a considerable technical challenge, and within the resulting system I have achieved all my major objectives. The method of disaggregating the total problem and the system into separate, modular, components worked very successfully and the system produced has a high level of integrity and reliability. The disaggregation of the database into isolated areas, holding name directories, source records and population structures, is an important design feature, which has also helped to disaggregate the total linkage problem into a series of more manageable sub-problems, capable of solution.

Finally, it is significant that the results of the whole linkage operation, as represented in the Population Database, are expressed in
a quite general way, in terms of persons, families and events, and so they are in no way constrained by the structural characteristics of the original source records which were presented for linkage. This is a crucial feature of the design, in that it means that the system possesses a high level of generality and flexibility, and that it has the potential for allowing entirely new types of source records to be incorporated and linked into the Population Database with minimal change required.

It should be clear that the system as presently implemented exhibits a considerable degree of generality and that its strategies are not heavily constrained by the particular record universe which was used. There is, however, a particular feature of the design where some refinement is possible, and it is to this that I shall now address my attention.

An examination of the detailed record matching and linking strategies described in this chapter will reveal that they are organised in a disaggregated rather than unified way. Thus, there is a separate part of the strategy for dealing with census records, another part for dealing with marriage records, and so on. While such an arrangement is the natural outcome from a research project employing an iterative and evolutionary development process, there are concomitant disadvantages, which are evidenced in the resulting system. Most prominently, in the present case, is the overlap which exists between the separate parts of the linkage strategy: this can lead to duplication of the individual 'rules' which are employed when
records are being matched. For example, for many nominal record types, there is the constraint which requires that only one occurrence of that record type may be associated with any one individual. Thus, an individual may have only one baptism, one burial and may appear only once in the 1851 census. Where the linkage strategy is disaggregated this particular rule needs to be duplicated wherever it is required. Such duplication requires the writing of more code and it can lead to inconsistent operation where the separate incarnations of the same rule are not applied in an identical fashion.

An important refinement to the overall linkage strategy would therefore entail the convergence of the separate parts of the strategy to produce a more integrated and flexible system. This would require some fundamental changes, and proposals for these are examined in the next section.

10.3.2 Proposed Modifications to the Linkage Strategy

The module which currently carries out the record matching operations is 'NWNAMES' (see Figure 10.1). Thus, for example, when it is necessary for a census occupant to be searched for in the Population Database, module 'STCELINK' invokes 'NWNAMES' and provides for it name and other information extracted from the census record. 'NWNAMES' then searches in the Population Database for a 'P-PERSON' record whose details sufficiently match the information provided. There is a similar interface between module 'UUMARLINK' and 'NWNAMES'
to provide corresponding searching and matching facilities for each of the marriage partners referred to in a marriage record.

Since census records and marriage records contain differing amounts of personal information, strategies appropriate to each kind of record must be provided: for the present implementation the corresponding strategies are described in sections 10.1.4 and 10.1.6. The strategies are implemented within 'NWNAMES' by separate sections of code, and there are additional sections of code appropriate to the matching strategies for other types of source records. In other words, the code within 'NWNAMES' is source record type-dependent, i.e. it 'needs to know about' the characteristics of the individual types of records. This is undesirable for two reasons. Firstly, from the system design viewpoint 'knowledge about' census records, for example, should be restricted as far as possible to the appropriate modules: 'STCELINK', 'SRSTORECEPOP' and so on. And secondly, from a practical viewpoint, as each new type of source record is introduced new sections of code must be introduced into 'NWNAMES'.

A significant refinement to the overall linkage strategy would therefore be achieved by redesigning module 'NWNAMES' so that it was completely source record type-independent. The resulting matching strategy would also then be completely source record type-independent. Let us consider how this might be achieved. Essentially a caller of 'NWNAMES', such as 'STCELINK', would be required to provide for 'NWNAMES' all the relevant identifying information about an individual which it has been possible to extract from a particular source record.
It is proposed that this information would be set up in a standardised 'person occurrence template' (38), and for any particular invocation of 'NWNAMEs' many of the item fields in this template would be empty, reflecting the fact that corresponding information was not available. The function of 'NWNAMEs' would then be to scan through all the relevant 'P-PERSON' records in the Population Database, compare the contents with the contents of the template and produce a score for each 'P-PERSON' based on the measure of agreement between the two. Where the highest score was greater than a particular preset 'cut-off' value, 'NWNAMEs' would report back to the caller that a matching 'P-PERSON' record had been found. A similar scheme could be provided for locating a matching 'P-FAMILY' record in the Population Database and a corresponding 'family occurrence template' would be provided.

One of the main difficulties in implementing such a scheme as this is the design of the person and family occurrence templates. These must have a sufficiently generalised structure so that they can cope with all the nominal record types which may subsequently need to be incorporated.

Let us consider first a person occurrence template. This must be capable of holding the essential identifying information about a person, as it might be extracted from a nominal record. The individual fields would be held in an appropriate internal coded form, with a special value used to indicate that the relevant information is missing. The following are the proposed contents:
- date of occurrence
- name information (at end of occurrence, where applicable (39))
- maiden surname
- female surname at beginning of occurrence
- gender
- marital status at start of occurrence
- marital status at end of occurrence (40)
- computed date of birth
- age indicator code, e.g. 'infant', 'minor', 'over 21'
- computed date of death
- death indicator code. When set this indicates that an explicit posthumous reference has been made to the person in a source record.
- birthplace information, including an indication of whether the birthplace was named explicitly or whether it has been implied from the residence in a baptism record
- illegitimacy indicator
- computed date of birth of a child of the individual (41)
- occurrence exclusion key. When set this indicates that this type of occurrence may exist only once for each 'P-PERSON' in the Population Database. The value of the key will indicate the type, e.g. 'baptism'. Each 'P-PERSON' record will have a field for each of the occurrence keys, and when set will indicate that the corresponding occurrence has already been linked.
- person exclusion fields. When set these can be used to identify people in the Population Database who are to be
excluded from the search (42). Such fields can be used to avoid the confusion which may arise where two people with the same name (e.g. a father and son) are referred to in the same record.

- owning family identification. This field identifies the person's family of origin, i.e. the owning 'P-FAMILY' record in the 'P-CHILD-SET' (43). It will only be set where the person's family of origin has already been located in the Population Database.

- owned family identification. This field identifies a marriage union for the individual, i.e. an owned 'P-FAMILY' record in either the 'P-FATHER-SET' or the 'P-MOTHER-SET'. It will only be set where such a marriage union has already been located in the Population Database.

In a similar fashion one can devise a family occurrence template. The following are the proposed contents:

- date of occurrence
- husband name information
- wife maiden name information
- wife surname at beginning of occurrence
- the following information for each partner and as specified for the person occurrence template:
  - marital status at start of occurrence
  - marital status at end of occurrence
  - computed date of birth
  - age indicator code
birthplace information
- computed date of birth of a child
- occurrence exclusion key. This would operate in the same way as the corresponding field in the person occurrence template.
- union start indicator. When set this indicates that the date of the occurrence is the date of the start of the marriage union.
- union end indicator. When set this indicates that the date of the occurrence is the date of the end of the marriage union (44).

Having designed the two templates it would then be necessary to draw up the two sets of rules for matching person occurrences with 'P-PERSON' records and family occurrences with 'P-FAMILY' records. Since it would now be possible to express these rules in a completely source type-independent form it is anticipated that the resulting overall strategy would now be considerably simpler to implement and refine. And this improvement could also represent a significant step forward in the attempt to provide more generalised record linkage mechanisms.

Let us consider what the source type-independent rules might look like, specifically in the case where person occurrences were being matched against target 'P-PERSON' records. The following represent some of the more fundamental tests which would need to be applied:
1. chronological tests

Is the date of the person occurrence within the life-span of the target 'P-PERSON'? Where the computed date of birth and/or computed date of death are present in the person occurrence, are these dates compatible with the known life-span of the target 'P-PERSON'? In the absence of such information, is there comparable date of birth compatibility when parents and/or children of the two individuals are taken into consideration? And most fundamentally, what is the discrepancy between the computed dates of birth for person occurrence and target 'P-PERSON'?

2. 'once-only' event test

If the person occurrence corresponds with an event which is a 'once-only' event in the life of the individual (e.g. presence in the 1851 census) then does the target 'P-PERSON' already 'own' such an occurrence? If so, then this would constitute a bar to linkage.

3. marital status test

Are the date and marital status information associated with the person occurrence consistent with the corresponding information for the target 'P-PERSON'? Thus, for example, if the person occurrence has marital status 'unmarried', does the
'P-PERSON' have a prior 'married' event?

Tests of a similar, source type-independent form could be applied when matching family occurrences with 'P-FAMILY' records.

Finally, I would like to propose an additional refinement to my overall strategy. A central element of my approach is that one should proceed by linking the most information-rich records first and progressively move to the information-weak ones. I therefore propose that each record in the Source Database should carry an information 'strength' indicator value. This would be computed by the Source Translation Subsystem on the basis of the information content of the record. At the time of record linkage it would then be possible for the system to order the linkage of the records strictly in accordance with their information 'strength' (45).
NOTES

1. A verification of this assumption is provided in Section 10.2.

2. A generalised approach has been adopted for the handling of all birth and baptism records, and this is reflected in the method of storing the information in the Source Database. The strategy for linking such records must, however, be content-dependent, and in the present implementation there are facilities for linking only post-1812 baptisms.

3. A similar comment can be made for death and burial records as was made for birth and baptism records (see Note 2). Thus, in the present implementation there are facilities for linking only post-1812 burials.

4. It will be observed that the characteristics of the linkage strategy outlined thus far mirror fairly accurately the informal manual methods of a genealogist. Given a broad range of record types he would tend to start with the most detailed and useful ones first and only move on to the others after establishing the main family groupings. Or, again, the approach is similar to that normally used when tackling a jigsaw puzzle. One begins with the 'easy' pieces, i.e. corners and edges, and progressively moves to the more 'difficult' ones, i.e. middle pieces with monochrome colouring and little visual content.

5. The user might, for example, wish to examine the comparative performances of the linkage strategies when the records are linked in different orders. Should such separate runs produce identical, or near identical, population structures then this would tend to confirm the validity of the linkages.

6. With earlier and more 'difficult' data, where ambiguities are considered to be rather more likely to occur, strategies for measuring 'correctness' must clearly be incorporated as an integral part of the actual linkage process.

7. The information needed to establish familial linkages may not always be explicitly present in the record. Thus, in 1841 census records 'relation to head' information is not generally provided, and for post-1841 census records familial relationships among those not related to the head are normally not specified. For such individuals tentative linkages could be established, based on the name, age and other information present. The required logic is, however, not included in the present strategy.

8. In a more elaborate implementation of this linkage strategy an attempt would be made to interrogate all the members of the 'P-COUPLE-NAM-SET', select the most likely 'P-FAMILY' candidate and only then decide whether it was a suitable match. However,
with the present database size there was usually only one 'P-FAMILY' record in each 'P-COUPLE-NAM-SET', and so the simple strategy has proved adequate. The amount of information available for each husband and wife in the census records, and specifically the date of birth and birthplace information, also guarantees that the present, simple strategy will only rarely create a false link.

9. Initially, only the first test was applied, but this failed to detect at least one 'true' match where an incorrect birthplace had apparently been entered. The inclusion of the second test permitted the birthplace error to be overlooked. It is conceivable that one would wish to refine the strategy further after testing the system with a larger dataset.

There has been considerable research into the precision of age statements in census records. See, for example, Anderson 1972, Razzell 1972, Tillott 1972, Kelly 1974 and Wrigley 1975. The general consensus is that age reporting is reliable, with most ages correct to within two years. However, Kelly's experience with American census data suggests that age reporting is less reliable there, and that age discrepancies of greater than four years are not uncommon (Kelly op cit, 76).

From these findings it is clear that a linkage strategy which permitted age discrepancies of only one year could risk the loss of a number of true links. Thus, the setting of the lower limit in my algorithm at two years seems appropriate. The danger in setting a moderately high upper limit (in my case, six years) is that I risk introducing false links. However, the fact that in this instance I am matching both the husband and wife names in combination should ensure that the risk is minimal.

Research into the precision of birthplace statements in census records suggests that there are significant inaccuracies (Anderson op cit, 75 and Wrigley op cit, 316). However, there is evidence that birthplace statements for natives of the parish of enumeration are very accurate (Razzell op cit, 131).

10. As for Note 8, in a more elaborate implementation of this linkage strategy I would aim to interrogate all the members of the 'P-BIRTH-NAME-SET', select the most likely 'P-PERSON' candidate and only then decide whether it was a suitable match. However, with the present database size there was on average less than two 'P-PERSON' records in each 'P-BIRTH-NAME-SET', and so the simple strategy has proved to be satisfactory.

11. The score values can be adjusted to reflect the relative importance which the individual fields are considered to merit. In the example provided 'matching gender' is given a high value, viz. 1000, thus emphasizing its importance. If, alternatively, it was decided that 'matching gender' should be accorded less significance then this could be achieved by reducing the associated score value. Winchester has described this type of matching algorithm as one based on 'simple subjective weights' (Winchester 1970, 120).
12. The age discrepancy settings used here are consistent with those used when matching a husband and wife in combination (see Note 9). In this case, however, a reduced upper limit (viz. four years) is adopted because of the increased risk of introducing false links when matching one person in isolation.

13. The connection of any type of event record will potentially provide new information which can be used to update the 'P-PERSON' record. In some situations, the addition of this new information will require that changes should be made in the way that the 'P-PERSON' record is connected to the Cohort and/or Name Indexes. For example, when a maiden name is dynamically inserted in a female 'P-PERSON' record it becomes possible for this record to be connected in the appropriate 'P-BIRTH-NAME-SET'.

14. The 'identity' in this context refers to the unique key value which IDMS automatically gives to a record and which is communicated to the user's program when the record is created or located. The program can make subsequent accesses to the record merely by submitting this key value to IDMS: such 'short cut' accesses will be much more efficient and convenient than repeated use of the normal CALC and set operations.

15. This use of the primary family identity is now thought to be potentially too restrictive. Where a 'P-FAMILY' record has previously been created merely to connect two or more siblings then this would clearly give a 'no match' when compared with the primary family. In practice this situation should occur very rarely since census records are by default linked in chronological order, and so the record containing the parents and family will normally be linked before a record containing only siblings. In this situation there would then be no requirement to create a special 'P-FAMILY' record to connect the siblings.

16. The connection here will not be automatic since the head may have married more than once and the 'P-FAMILY' record which he owns in the 'P-FATHER-SET' may not relate to the family of origin of the son in the current household. In such a situation a limited amount of checking must be carried out to ascertain whether the 'P-FAMILY' record is the required one. If it is deemed not to be then it will be necessary to create an additional 'P-FAMILY' record to provide the 'correct' head-son connection.

17. In view of the unlikelihood of this particular situation arising the logic needed to accomplish the merging of the 'P-FAMILY' records has not been installed.

18. The same reservations are now felt about the use of the secondary family identity in this test as were expressed about the use of the primary family identity in Note 15.

19. Strictly one would expect to find no child born before the marriage date. This 2-year safety margin is included so that a genuine 'late' marriage may not be excluded and also to cover any
small errors in reporting dates and in date of birth estimations.

20. It is this double connection to the Name Index which enables the bride's pre-marriage and post-marriage events to be linked to her 'P-PERSON' record.

21. For reasons of simplicity the current implementation strategy does not make use of the additional information normally present in a post-1837 marriage which serves to identify the father of each marriage partner.

22. The 3-year tolerance included here will allow for some deliberate age misrepresentation and/or error in the computed date of birth information.

23. The 2-year tolerance included here will allow for some error in the computed date of birth information. It does, however, seem unlikely that there would be misrepresentation when 'minor' was specified, and so there is no justification for including additional tolerance in the age comparison.

24. This criterion is intended to distinguish a parent and child who have identical names.

25. As was pointed out in Section 10.1.4 it is occasionally necessary to create a 'parentless' 'P-FAMILY' record to connect a number of siblings who have been located in a census household. And it is just such a 'P-FAMILY' record which at some stage in the linkage process may have to be merged with another 'P-FAMILY' record which has been independently created to link the husband and wife.

26. The 3-year tolerance included here will allow for some error in the recording of the two sets of date information, and it will additionally allow for the situation where a father is posthumously named in the baptism record of a child born near (and perhaps just after) his death.

27. See Note 26. Since a child can be born after the death of its father but not after the death of its mother the 3-year tolerance may in this case be slightly over-generous.

28. While there may be occasional exceptions to this rule it was considered that the effort needed to implement a more comprehensive search strategy was not justified.

29. Record destruction is required when two 'P-PERSON' or two 'P-FAMILY' records are merged. Consider, for example, the process of merging two 'P-PERSON' records. This first requires that one of them is nominated to be the 'survivor'. Information from the second one is then copied into the survivor and is itself destroyed.

    The process of merging involves some additional rather complex activities. For example, where the 'P-PERSON' record to be destroyed has one or more event records connected in the
'P-LIFE-EVENT-SET' each of these needs to be disconnected from this set and reconnected into the corresponding set for the surviving 'P-PERSON'. In a similar fashion the familial connections, in the 'P-FATHER-SET', 'P-MOTHER-SET' and 'P-CHILD-SET', must also be disconnected and reconnected, where necessary, to the survivor 'P-PERSON'.

30. See Section 10.1.1.

31. Each of these sources was drawn from the 1851 Census National Sample of Great Britain. They were randomly selected from the largest 'clusters' (i.e. communities) in the Sample. The five communities are Bucklebury, Berkshire (PRO H.O. 107/1691), Grimstone, Norfolk (/1828), Dodbrooke, Devon (/1876), Deeping Fen, Lincolnshire (/2095) and North Sunderland, Northumberland (/2420).

32. Each community was considered to be a base zone. Any birthplace within it was allocated the code 'B', all other birthplaces were allocated the code 'N'. This is consistent with the zoning scheme described in Section 3.2.2.

33. See Note 9.

34. This should be valid where each ambiguity involves two people only. For my data there was only one exception to this, viz. an ambiguity involving three people.

35. See Chapter 3, Note 1.

36. See Section 10.1.1.

37. For example, for very small populations there will be negligible duplicate forename-surname combinations, and so there will be virtually no possibility of nominal ambiguity. As the population is increased it may be that a critical population size is eventually reached, above which the proportion of duplicate forename-surname combinations in the population is sufficient to cause a sharp increase in the incidence of nominal ambiguity problems. In the absence of appropriate empirical data it is difficult to quantify such an effect.

38. Macfarlane has explored the use of a similar scheme (Macfarlane et al 1983, 29).

39. For a female at a marriage ceremony there is normally a change of surname, and so space must be provided for potentially three surnames: her maiden surname, her pre-marriage surname and her post-marriage surname.

40. For both partners at a marriage ceremony there will normally be a change of marital status, and so the pre-marriage status and post-marriage status must both be recorded.
41. When searching for a parent referred to in a baptism record no age information will normally be present for the parent and so the computed date of birth of the child can be of value in narrowing the search.

42. The identification could, for example, either be in the form of the unique six character 'P-PERSON' identifier or else, and perhaps more conveniently, the unique key value used by IDMS (described in Note 14).

43. There are two distinct methods for uniquely identifying a 'P-FAMILY' record: these parallel the methods for uniquely identifying a 'P-PERSON' record, described in Note 42.

44. This indicator would be set, for example, where a burial record was encountered which named a deceased husband and an extant wife. Strictly in this case it would be the computed date of death of the husband, rather than the date of the burial event, which would provide the date of the end of the marriage union.

45. A further development of this proposal would be desirable for certain kinds of records. Thus, for example, census household records are essentially 'hybrid' records, typically containing substantial information about the head's family, together with relatively meagre information about the remaining occupants. There would, therefore, be a sound argument for disaggregating these into separate entities in the Source Database, each with its own strength indicator value. As a result, at record linkage time the information-rich section of the household, i.e. the head's family, would be linked at an earlier stage in the process than the remainder.
In the last chapter I examined the provision of facilities for linking and organising nominal records in the Population Database. What I shall now consider is the nature of the additional facilities which are needed to retrieve information from the Population Database in an appropriate form.

So far, I have chosen to view the major stages in the total record linkage process as providing progressive transformations of the nominal data. In the earlier stages the individual records were transformed from a user-oriented to a system-oriented form. In the later stages the records were then subjected to radical restructuring, involving the disaggregation and re-aggregation of individual records. As it ultimately resides in the Population Database the nominal data is organised into potentially very complicated network structures, modelling, as it does, the equally complicated 'real world' genealogical, residential and life-history patterns.

What I shall now address attention to in the present chapter is the final transformation of the nominal data. The objective of this transformation will be to 'complete the circle', i.e. by finally restoring the data to a user-oriented form. This process is conceptually the most difficult, since the structures which are held in the Population Database are potentially very complex, and in order
to satisfy a user's access requirements, it will be necessary for such structures to be 'collapsed' and reduced into a wide range of simpler forms.

At an elementary level of analysis the user may merely wish to obtain some simple statistical information about the contents of the Population Database, such as the total population, the sex-ratio for particular cohorts, the proportion of the population in various occupational groups, etc. At a more complex level he may wish to relate information which is held in dispersed parts of the database. He could, for example, be interested in relating characteristics of the household size in the 1851 census to the residential and occupational mobility of the head of each household. Finally, for a user with a particular interest in family structure there may be a requirement for some means of having genealogical structures displayed graphically, in the form, perhaps, of family trees.

It has not been feasible, within the scope of the present research and of the particular database management system which was employed, to satisfy such a wide diversity of analytical and display requirements. The main thrust of the research has been directed towards the development of the linkage strategies themselves, and the provision of facilities for subsequently manipulating the linked data has been accorded a secondary significance. Indeed, the preferred option would be to have the provision of such facilities completely removed from the responsibility of the linkage system itself. Thus, an appropriate strategy would be to arrange for the transfer of the
entire contents of the Population Database to a suitable relational database system, which would then be able to provide a wide range of analytical and display functions. I shall explore this proposal in more detail in Section 11.4.

Throughout the remainder of this chapter I shall examine the range of analysis facilities which have been incorporated in the present system. The program which has responsibility for providing these facilities is the Population Analysis Subsystem, and in Section 6.6 I proposed that the prime function of this subsystem, in its initial implementation, should be to act as a monitoring device for the record linkage operations. As such, it should be designed to display details about selected groups of people and families in the database. A comparison of these details with a display of the original source records should then enable the 'correctness' of the linkage to be assessed.

A range of facilities for assisting this monitoring process was described in Section 6.6. Let us recall briefly the way in which they are intended to be used. Typically, a user would submit a particular surname, and the subsystem would proceed to locate each person and/or family with that surname, or a synonym of it, and arrange to retrieve and display the associated information from the database (1). In a similar fashion, he would submit the same surname to the Record Linkage Subsystem and request it to display the associated source records for that name (2). These two sets of information, giving the results of the linkage and also the records which contributed to it,
when examined side-by-side, should enable the linkage to be readily validated (3).

Although this monitoring role is regarded as being the most important one it was proposed in Section 6.6 that other, potentially useful methods of examining the linked data should also be provided. Thus, it was suggested that the user should be able to obtain displays of subsets of the persons and families in the Population Database which share a variety of criteria. For example, it should be possible to have displayed all the males born in the base zone in a specified range of years. Or again, it should be possible to have displayed all those who had a particular occupation. Provided with such displays one should then be in a position to embark on some limited substantive analyses of the population data: the preliminary results of such analyses for the Elwick Hall parish data are examined in the next chapter.

In my examination of the problems of displaying information from the Population Database I shall, for convenience, subdivide the analysis into two distinct parts. In the first part I shall investigate the problems of displaying information for each 'case', i.e. each person or family in the database in which the user is interested. And in the second part I shall then examine the way in which the Cohort, Name and Occupation Indexes may be used to select and gain access to the individual 'cases'. Later, I shall observe the way in which these access and display facilities are implemented, by examining the internal design of the Population Analysis Subsystem.
Finally, at the end of the chapter, I shall consider the contribution which a relational database system could make towards the provision of improved analytical facilities.

11.1 POPULATION INFORMATION DISPLAY

In Section 6.6 I proposed that the user should be able to select the amount of information which he requires to have displayed for each person or family in which he is interested. And I described the way in which he could use the 'DETAILLEVEL' command to specify his selection. Thus, for example, if he were to submit the following command:

```
PP/DETAILLEVEL/PERSONS=MAXIMUM/FAMILIES=MAXIMUM
```

then the system would subsequently arrange to display comprehensive information about each person and family. For example, for each person it would display the name, gender, date of birth, birthplace and the person's unique identification key, together also with details of his parents, marriage partner(s), number of children and the individual 'events' in his life.

I shall now examine the processes which are involved in providing such displays. In the next section I shall consider how information about a person may be displayed, and for this I shall assume that by
some means a particular 'P-PERSON' record has been selected and located. I shall also assume that 'maximum' details have been requested. In the following section I shall then carry out a corresponding analysis of the processes which are involved in displaying information about a family.

11.1.1 Person Information Display

An example of the way in which information about a person is displayed was provided in Figure 9.6. Essentially this display, illustrating the life-history of Frances Wilson, is made up of four separate sections, as follows:

1. an initial line containing the person's unique identification key, name, gender, computed date of birth and birthplace. Where the person is known to be illegitimate the character 'I' will precede the gender code.

2. information about each parent, where known. The parent's unique identification key, name, computed date of birth and birthplace are provided (4).

3. information about each marriage. The unique identification key for the relevant family is displayed, together with information about the spouse, with the same details being displayed as for each parent. In addition, there are counts
of the numbers of sons and daughters.

4. information about each event in the person's life, presented in chronological order. The year and type of event are displayed first, followed by more specific details obtained from the corresponding source record. The precise date of the event, together with the appropriate source record unique identification key are always present. For each census event there are additionally the usual personal details: relation to head, marital status, occupation, computed date of birth and birthplace, together with the enumeration district and household address information. For each vital event there are additionally the place of registration and the religious denomination (where applicable), together with the particular personal details which were originally submitted or have been derived: marital status, computed date of birth, birthplace, residence, occupation, etc., as appropriate.

Let us consider now the nature of the processes which are needed to produce such a life-history display. The information which is displayed is essentially contained in only six of the record types in the Population Database: these and their associated sets are illustrated in Figure 11.1.

I shall assume for the purposes of this analysis that a particular 'P-PERSON' record has been nominated for display and has been located. In order for the first section of the display to be
Figure 11.1 The Population Database: person and family information display
(* = CALC record)
produced it is, in fact, necessary only for fields within this 'P-PERSON' record itself to be accessed. Prior to display, however, all the relevant internal codes need to be converted back into strings of characters. The processes which are involved are precisely the same as those described in Section 9.1.1 for the corresponding display of information from the Source Database.

In order to produce the second section of the display, viz. that relating to the parents, it is necessary to move to other 'P-PERSON' records in the database to locate the required information. Starting from the nominated 'P-PERSON' record it is first necessary to locate the owning 'P-FAMILY' record in the 'P-CHILD-SET': this represents the person's family of origin. If it exists and it also has connected an owning 'P-PERSON' record in the 'P-FATHER-SET' then this record can be retrieved, and information about the father can be located and displayed. In a similar fashion if there is an owning 'P-PERSON' record in the 'P-MOTHER-SET' then this can be retrieved, and information about the mother can be located and displayed.

For the third section of the display, viz. that relating to marriages, it is necessary to start again from the nominated 'P-PERSON' record and attempt to locate one or more 'P-FAMILY' records in either the 'P-FATHER-SET' (for a male) or the 'P-MOTHER-SET' (for a female). These will represent the person's families of procreation. For each 'P-FAMILY' record thus located and retrieved the relevant information about the family and number of offspring may be obtained and displayed. Where the nominated person is male an attempt can also
be made to locate the owner of the 'P-FAMILY' record in the 'P-MOTHER-SET'. If the required 'P-PERSON' record exists then it can be retrieved, and details of the wife can be obtained and displayed. Where the nominated person is female a similar process can be employed to locate and display details of the husband.

In order to produce the final section of the display, that relating to events, it is necessary to retrieve each member record in the 'P-LIFE-EVENT-SET' for the nominated 'P-PERSON' record. For each 'P-EVENT' record located it is possible to obtain and display information about one vital event in the life of the individual who was nominated. Likewise for each 'P-CENS-EVENT' record located it is possible to obtain and display information about one census event. In this case, however, it will also be necessary to retrieve the owning 'P-CENS-HOUSEHOLD' record in the 'P-CENS-OCC-SET' and then the owning 'P-CENS-ENTRY' record in the 'P-CENS-HSE-SET' in order that the associated household and enumeration district address information may be obtained.

The information described above is regarded as being the notional 'maximum' which is likely to be required about any one individual (5). The notional 'minimum' is restricted to the information contained in the first section only, i.e. that which can be obtained from the nominated 'P-PERSON' record itself. The 'DETAILLEVEL' command, described earlier, can be used to request that information is to be provided at the maximum, minimum or some intermediate level of detail. At the intermediate levels information in the second, third and fourth
sections may be optionally displayed or suppressed.

11.1.2 Family Information Display

The end product of family-based record linkage is not merely a set of personal dossiers, but also a corresponding set of family dossiers, holding the 'life events' and other details of individual families. In this section I shall examine the form and content of such family dossiers.

An example of a family dossier is provided in Figure 11.2. This display illustrates the family of Robert and Frances Pattison (6). The method of presentation precisely parallels that described in the previous section for information about a person. In this case, however, the display is made up of only three separate sections, as follows:

1. initial, brief details about the family and the associated marriage event, where this has been located. The details include the family's unique identification key and the names and birthplace zone codes of the father and mother. Additional information from the marriage record includes the source record unique identification key, the date and place of the marriage, the type of ceremony (i.e. whether by banns or by licence) and for bride and groom the marital status and the signature/'X' indication.
Figure 11.2  The displaying of family information by the Population Analysis Subsystem:
the family of Robert and Frances Pattison
2. more detailed information about each parent, where available. The parent's unique identification key, computed date of birth and birthplace are provided.

3. information about each child of the marriage, presented in correct family order. For each child the unique identification key is provided, together with the name, gender, computed date of birth and birthplace.

Let us consider the nature of the processes which are needed to produce such a display. The relevant records and sets which must be accessed are a subset of those shown in Figure 11.1. Indeed, the only entities involved are the 'P-PERSON' and 'P-FAMILY' records and the three interconnecting sets.

Once again, as for the previous life-history analysis, I shall assume that a particular record, in this case a 'P-FAMILY' record, has been nominated for display. In order for the first section of the display to be produced it is, in fact, necessary only for the fields within this 'P-FAMILY' record itself to be accessed. And, just as for the 'P-PERSON' record display, all the relevant internal codes need to be converted back into strings of characters prior to display.

In order to produce the second section of the display, viz. that relating to the additional information about the parents, it is necessary to move to other records in the database to locate the required information. Starting from the nominated 'P-FAMILY' record
it is necessary to locate the owning 'P-PERSON' record in the 'P-FATHER-SET' and the owning 'P-PERSON' record in the 'P-MOTHER-SET'. From these records the additional information about the father and mother, respectively, can be located and displayed.

For the final section of the display, viz. that relating to the children, it is necessary to retrieve each member record in the 'P-CHILD-SET' for the nominated 'P-FAMILY' record. Each 'P-PERSON' record thus located provides the information about one child of the marriage.

The above information is regarded as being the notional 'maximum' which is likely to be required about a particular family. The notional 'minimum' is restricted to the information contained in the first section only, i.e. that which can be obtained from the nominated 'P-FAMILY' record itself. Once again, as for person information, the 'DETAILLEVEL' command can be used to request that the family information is to be provided at the maximum, minimum or some intermediate level of detail. At the intermediate levels information in the second and third sections may be optionally displayed or suppressed.
11.1.3 Statistical Information Display

It is necessary finally to consider the arrangements for displaying brief statistical information about the contents of the Population Database, such as the number of persons and families present. Typical output was illustrated at the end of Chapter 6.

The processes which are needed to display this information are simple. It is necessary merely for the five special 'P-ITEM-CNT' records to be retrieved from the Population Database in turn: from each a count value is obtained and displayed.

11.2 INDEXING PERSONS AND FAMILIES

Having examined the nature of the processes which are needed to display information about nominated persons and families I shall now consider how particular subsets of persons and families may be located prior to display. The types of selection commands which are provided for the user were described in Section 6.6. I shall now demonstrate how the various selections may be accomplished by appropriate accesses to the indexes which were constructed during the earlier record linkage operations. I shall consider, in turn, accesses via the Cohort, Name and Occupation Indexes (7).
Access via the Cohort Index

A description of the Cohort Index was provided in Section 9.2.2, and in the last chapter I examined the processes by which individual 'P-PERSON' and 'P-FAMILY' records are connected to it during record linkage. What I shall consider in this section is the reverse process, i.e. how the index may be used to gain access to particular groups of persons and families.

I shall consider first the access to persons, and shall examine the nature of the processes which are needed to provide a response to the following two typical cohort selection commands:

\[
\begin{align*}
PP/\text{SUBSET=PERSONS/COHORT=1831-1840/B/M} \\
PP/\text{SUBSET=PERSONS/COHORT=ALL/N/F}
\end{align*}
\]

The first command identifies the group of males born in the base zone in the period 1831-40, while the second identifies all females not born in the base zone.

The relevant records and sets which are used in both the cohort and name indexing selections are shown in Figure 11.3. In order to respond to the first command it will be necessary to locate ten 'P-PERSON-COHORT' records, one each for the years 1831-1840. For each of these 'P-PERSON-COHORT' records one must then arrange to retrieve each 'P-PERSON' record which is connected to it in the 'P-PERSON-SET'.
Figure 11.3 The Population Database: access to persons and families via the Cohort and Name Indexes (* = CALC record)
Thus, to locate the first 'P-PERSON-COHORT' record it will be necessary to submit the appropriate CALC code to IDMS. This code will be constructed from the codes 'M' and 'B' and the string '1831', i.e. the combined code will be 'MB1831'. Having located and retrieved the corresponding 'P-PERSON-COHORT' record all that remains is to retrieve each 'P-PERSON' member record in the 'P-PERSON-SET'. There will be one such record for each male born in the base zone in the year 1831. This process is then repeated for each of the other 'P-PERSON-COHORT' records in the range of years specified.

In order to respond to the second command initial access must be made to the upper level of the Cohort Index. The record which must first be located is the 'P-ALL-PERSONS' record for the female natives of the non-base zone. As before an appropriate CALC code must be submitted to IDMS: in this case the code will be constructed from the codes 'F' and 'N', i.e. the combined code will be 'FN'. Having located and retrieved the corresponding 'P-ALL-PERSONS' record the next step is to retrieve in turn each member 'P-PERSON-COHORT' record in the 'P-PER-COHORT-SET'. And as each 'P-PERSON-COHORT' record is retrieved its 'P-PERSON' member records in the associated 'P-PERSON-SET' are themselves retrieved in turn. In this way access to all the persons in the major subgroup may be achieved.

Access to families via the Cohort Index can be provided in a similar fashion. However, in view of the resource constraints of the project the required facilities have not been included in the present implementation. And bearing in mind the size of the test Population
Database they would have been of minimal usefulness (8).

11.2.2 Access via the Name Index

A description of the Name Index was provided in Section 9.2.3, and the relevant records and sets which are involved in its operation are shown in Figure 11.3. Since the Name Index is used extensively during record linkage operations, I have already, in the last chapter, had an opportunity to examine certain aspects of its method of use. I shall now examine the ways in which it is used during population analysis operations.

Let us consider first the access to persons and the nature of the processes which are needed to provide a response to the following two typical name selection commands:

\[
\begin{align*}
\text{PP/SUBSET=PERSONS/NAME=JOHN SMITH} \\
\text{PP/SUBSET=PERSONS/SURNAME=SMITH}
\end{align*}
\]

The first command identifies all people having the name 'JOHN SMITH' or a synonym of it, while the second identifies all people having the surname 'SMITH' or a synonym.

In order to respond to the first command the initial task is to attempt to locate the 'P-PERSON-NAME' record for 'JOHN SMITH'. The method of doing this was described in Section 9.2.3. All that needs
to be added here is that in the present situation there must be facilities available to provide access to the Christian Name and Surname Directories so that the internal code values for 'JOHN' and 'SMITH' may be obtained. Their corresponding major codes, used in combination, then provide the CALC code which is submitted to IDMS. Once the required 'P-PERSON-NAME' record has been located and retrieved the 'P-PERSON' member records in the 'P-BIRTH-NAME-SET' can then be retrieved. There will be one such record for each person called 'JOHN SMITH'.

In order to respond to the second command initial access must be made to the upper level of the Name Index. The record which must first be located is the 'P-SURNAME' record for 'SMITH'. Again, the precise method of doing this was described in Section 9.2.3. Having located and retrieved the required 'P-SURNAME' record the next step is to retrieve in turn each member 'P-PERSON-NAME' record in the 'P-PER-NAME-SET'. And as each 'P-PERSON-NAME' record is retrieved its 'P-PERSON' member records in the associated 'P-BIRTH-NAME-SET' are themselves retrieved in turn. Additionally, for a female 'P-PERSON-NAME' record there may be 'P-FAMILY' member records in the associated 'P-WIFE-NAME-SET'. For each of these it will be necessary to locate the 'P-FAMILY' record and then retrieve its owning 'P-PERSON' record in the 'P-MOTHER-SET'. In this way access to all the persons called 'SMITH' will be achieved: this will include females for whom 'SMITH' is the maiden or a married surname.
It was noted in Section 9.2.3 that each 'P-SURNAME' record contains the corresponding character string for the major surname and that each 'P-PERSON-NAME' record contains the appropriate character string for the major Christian name. These strings are expressly included so that they can be used in any display which is produced via the Name Index. Thus, for example, in Figure 6.3 the heading line 'WILLIAM BONE', which is underlined with asterisks, is constructed from such major name strings. The names which are displayed in the individual entries which follow are those which correspond with the codes obtained from the actual 'P-PERSON' records involved, and they will not necessarily consist of major name strings. It was considered that the deliberate use of major name strings in the headings helps to give the displays a tidy and consistent appearance. A similar policy is adopted for any display which is produced via the Occupation Index.

Next, I shall consider access to families via the Name Index, and shall examine the necessary operations which must be carried out in response to the following two commands:

\[
\begin{align*}
\text{PP/SUBSET=FAMILIES/COUPLE=JOHN,ANN SMITH} \\
\text{PP/SUBSET=FAMILIES/SURNAME=SMITH}
\end{align*}
\]

The first command identifies those families for which the marriage partners are called 'JOHN & ANN SMITH', while the second identifies all the 'SMITH' families. As usual, synonyms of the specified names are to be accepted as valid alternatives.
In order to respond to the first command the initial task is to attempt to locate the 'P-FAMILY-NAME' record for 'JOHN & ANN SMITH'. Again, the method of doing this was described in Section 9.2.3. Once the required record has been located and retrieved the 'P-FAMILY' member records in the 'P-COUPLE-NAM-SET' can then be retrieved. There will be one such record for each family whose marriage partners have the required names.

In order to respond to the second command initial access must be made to the upper level of the Name Index, to locate the 'P-SURNAME' record for 'SMITH'. This operation is precisely the same as for the initial part of the search for all persons having the surname 'SMITH'. Having located and retrieved the required 'P-SURNAME' record the next step is to retrieve in turn each member 'P-FAMILY-NAME' record in the 'P-FAM-NAME-SET'. And as each 'P-FAMILY-NAME' record is retrieved its 'P-FAMILY' member records in the associated 'P-COUPLE-NAM-SET' are themselves retrieved in turn. In this way access to all the 'SMITH' families will be obtained.

11.2.3 Access via the Occupation Index

A description of the Occupation Index was provided in Section 9.2.5, and the relevant records and sets which are involved in its operation are shown in Figure 11.4. It should be noted that, unlike the two previous indexes, the Occupation Index provides direct access only to persons. In the last chapter I examined the processes by
Figure 11.4  The Population Database: access to persons via the Occupation Index
(* = CALC record)
which individual 'P-PERSON' records are connected to the Occupation
Index via their 'event' records during record linkage. I shall now
observe the reverse process, i.e. how the index may be used to gain
access to particular groups of persons.

Let us consider the nature of the processes which are needed to
provide a response to the following two typical occupation selection
commands:

PP/SUBSET=PERSONS/JOBNAME=SHOEMAKER/C5
PP/SUBSET=PERSONS/JOBNAME=SHOEMAKER

The first command identifies all those people who had the occupation
'SHOEMAKER' in the 1851 census, while the second identifies all those
who had the occupation 'SHOEMAKER' at some point in their lives. As
usual, synonyms of the specified occupation are to be accepted as
valid alternatives.

In order to respond to the first command it will be necessary to
attempt to locate the 'P-OCC-CENS-ENTRY' record for the required
occupation-census combination. The method of doing this was described
in Section 9.2.5. Once this record has been located and retrieved the
'P-CENS-EVENT' member records in the 'P-OCC-CENS-SET' can then be
located. Finally, for each of these 'P-CENS-EVENT' records the owning
'P-PERSON' record in the 'P-LIFE-EVENT-SET' can be retrieved. Each of
these will be a member of the required group, i.e. those who had the
occupation 'SHOEMAKER' in the 1851 census.
In order to respond to the second command, i.e. for the more global search, it will be necessary to attempt to locate the 'P-OCC-ENTRY' record for the required occupation. Again, the method of doing this was described in Section 9.2.5. Having located and retrieved the 'P-OCC-ENTRY' record the next step is to retrieve in turn each member 'P-EVENT' and 'P-CENS-EVENT' record in the 'P-OCC-SET'. For each of these event records access to the associated person can be achieved by retrieving the owning 'P-PERSON' record in the 'P-LIFE-EVENT-SET'. However, since there can be several access routes between any pair of 'P-OCC-ENTRY' and 'P-PERSON' records (9), it is necessary to employ a strategy which will avoid locating the same 'P-PERSON' record more than once. In the present implementation the use of a special marker value, which is placed in only one event record for each person for a given occupation, enables this problem to be avoided. Thus, only for an event record which contains the special marker value is the owning 'P-PERSON' record in the 'P-LIFE-EVENT-SET' retrieved. Each of the persons thus identified will be a member of the required group, i.e. those who had the occupation 'SHOE MAKER' at some point in their lives.
11.3 THE CONTROL STRUCTURE OF THE POPULATION ANALYSIS SUBSYSTEM

I shall now briefly examine the internal design and control structure of the Population Analysis Subsystem. This is illustrated in Figure 11.5. As usual, for simplicity, the interfaces to the Input-Output Subsystem have been omitted.

The overall arrangement of the modules shown in Figure 11.5 has much in common with the corresponding one for those modules in the Record Linkage Subsystem which are responsible for selecting and displaying records from the Source Database. This arrangement was shown in Figure 9.4. The observed, structural similarity may be attributed to the similarity of their functions. Thus, both the Record Linkage and Population Analysis Subsystems are required to retrieve records from the database, convert their constituent fields from internal to external form, and then display them. And they are both required to carry out these functions in response to user-supplied commands. An obvious point of similarity in the structural organisation is the preponderence of database-handling modules: thus, of the 14 modules shown in Figure 11.5, 11 interface directly with IDMS. Once again, this reflects the functional bias of the subsystem. It has relatively little interaction with the user, but considerable interaction with both the Population and Directory Databases.

I shall now examine the flow of control through the subsystem and the main functions of each of its modules. Entry to the subsystem is
Figure 11.5 The Population Analysis Subsystem: structure and flow of control
via the top module in the diagram, 'MSPOPANALYSIS'. The module directly below this, 'P5ENTRYAN', has the usual initialising and terminating functions. Thus, it is responsible for initiating the opening of the input and output files and the database, and correspondingly at the end of the job for the closing of input and output files and the database. It also initiates the printing of opening and closing messages to the user. Finally, it has responsibility for processing the user's commands (10) and for forwarding appropriate action requests to module 'PNDBPROCESS'. The details of the particular action required are contained in an 'action control block', a large structure which is passed as a single parameter in the COBOL 'CALL' statement. This block contains a complete definition of the action required: the type of action (e.g. print statistics), whether person or family subsets are to be located, which index is to be used, as appropriate, the detail level of the displays, etc.

Module 'PNDBPROCESS' is responsible for supervising the opening of the database and for switching to the appropriate action processing modules. It calls 'TYCOHORTS' where a subset of 'P-PERSON' records is to be located via the Cohort Index, 'JYJOBS' where a subset of 'P-PERSON' records is to be located via the Occupation Index and 'NVNAMES' where a subset of 'P-PERSON' or 'P-FAMILY' records is to be located via the Name Index. Finally, it calls 'XZSTATS' where the brief statistical information about the contents of the Population Database is to be displayed. The action control block is passed down as a parameter with each of these calls so that the lower-level
modules are provided with the information which will enable them to carry out the required actions.

Where 'TYCOHORTS' is invoked to locate a subset of 'P-PERSON' records it first obtains its controlling parameters from the action control block: such parameters will indicate, for example, the range of years over which it is to search, the gender of the required 'P-PERSON' records, etc. It then proceeds to retrieve one or more records from the Cohort Index and to traverse the associated sets. For each 'P-PERSON' record in the subset which it locates it invokes module 'PLDBPRINT' to provide the appropriate information display.

Module 'JYJOBS' carries out parallel operations for the location of 'P-PERSON' records via the Occupation Index. And, again, module 'PLDBPRINT' is invoked to provide the appropriate information display for each 'P-PERSON' record which is located.

Module 'NVNAMES' operates in a similar fashion in relation to the Name Index, although, of course, it is also required to locate subsets of 'P-FAMILY' records. For each such 'P-FAMILY' record which it locates it invokes module 'PKDBPRINT' to provide the appropriate information display.

In order to display the required information for a person, module 'PLDBPRINT' first references the action control block to determine what level of detail is required. Where 'minimum' detail has been specified it will subsequently need to access only the 'P-PERSON'
record which has been nominated. Where more detail is required
'PLDBPRINT' will need to locate and retrieve other records in the
database. The overall logic for this was described in Section 11.1.1.
Module 'PKDBPRINT' is responsible for the comparable facilities for
family information display: the required logic was described in
Section 11.1.2.

All accesses to the name directories are handled by modules
'DQCONTRACT', 'DRDBSEARCH' and 'DSDBSEARCH'. 'DQCONTRACT' has overall
control and it invokes the other two modules to carry out the various
name and code conversions. 'DRDBSEARCH' arranges for a name string to
be converted into an internal code value by accessing the appropriate
name directory: such conversions are needed in the handling of the
user's commands. And 'DSDBSEARCH' arranges for the reverse
conversion, i.e. from an internal code value back into a name string.
This module is used heavily during the displaying of person and family
information. In a similar fashion module 'T3GETDATE' arranges for the
conversion of a computed date into a calendar date. This module
provides the same function for the Record Linkage Subsystem.

Finally, 'PMDBCLOSE' is the database closing module, which is
called both at the end of a normal run and also where an irrecoverable
error occurs in some database operation.
11.4 POPULATION INFORMATION DISPLAY VIA A RELATIONAL TRANSFORMATION

When the question concerning the selection of a suitable database management system for record linkage work was raised in Section 5.1.2 the central concern was with how well a particular system would aid the operation of the linkage process itself. The decision made then, viz. to use a CODASYL-type system, is still considered to have been the correct one when suitability for record linkage is regarded as the major criterion.

Where a CODASYL-type system proves to be less adequate is in its capacity for handling the post-linkage, analytic requirements, i.e. the kinds of requirements which the Population Analysis Subsystem has been designed to satisfy. Although this subsystem, as implemented, provides a wide range of indexing, subsetting and display facilities, it still does not have the level of retrieval flexibility which would ideally be desirable. Given that the Population Database contains a mass of inter-related data it would in fact be preferable to be able to examine and retrieve the data in accordance with one's own subsetting requirements and user view. With a CODASYL-type system such retrievals can only be provided by making continual enhancements to the database accessing programs.

To provide the desired flexibility a more appropriate solution would be to make use of the relational approach. In an enhanced version of the current system one might therefore choose to use IDMS to carry out the linkage operations, i.e. as at present, and then
provide a vestigial Population Analysis Subsystem, whose sole function
would be to off-load the contents of the Population Database into a
series of files which would then be transferred to a relational
database system. All subsequent analysis would then be carried out
using the relational Population Database.

Let us consider the processes which would be involved in
off-loading the data in a form suitable for transferral to a
relational database system. And, initially, I shall restrict
attention to the parts of the Population Database which model the
family structures, i.e. the 'P-PERSON' and 'P-FAMILY' records,
together with the three inter-connecting sets. In order to load these
family structures into a relational database it would be necessary to
replace them by two flat files \( (11) \), one containing person records and
the other containing family records. Since all 'P-PERSON' records
have an identical size and structure they are already in a suitable
form for easy off-loading into a flat file. In the actual off-loading
operation it would, of course, be necessary to gain access to every
'P-PERSON' record in the database: this could be readily achieved by
using the Cohort Index \( (12) \). Similar considerations apply to the
off-loading of the 'P-FAMILY' records.

Let us now consider the person-family connections, which in the
IDMS implementation are represented by set connections between the
records. In a relational system there is no set concept: instead, the
inter-record connections are provided by the presence of key values
which are stored in the records. Every record of a particular type is
given its own unique key value which serves to identify it. For example, suppose that a particular record 'A' has been given a key value of '27'. Then records 'B' and 'C' can be 'connected' to 'A' simply by each having the value '27' stored in an appropriate field within them.

Since in my IDMS implementation both 'P-PERSON' and 'P-FAMILY' records already have unique identification keys, e.g. 'P00338' and 'F00053', they are ideally designed for transferral to a relational database system (13). Let us consider then how some typical family connections could be represented by means of an arrangement of two flat files. I shall use as an example the family of Robert and Frances Pattison, which was illustrated in Figure 11.2. When transposed into the two flat files the organisation of the person and family records will be as shown in Figure 11.6. It should be noted that for illustrative purposes only the more significant data items have been included. Thus, for each person there is the unique person key, name, gender, date of birth and, where known, the family key for the person's family of origin (14). And for each family there is the unique family key, names of the marriage partners (15), date of marriage and the person keys for the father and mother (16).

An examination of this scheme will reveal that it is equivalent to the record and set organisation used in the current IDMS implementation. Consider first the eight children of the family. For each of these the family of origin field contains the key value 'F00053', and this links them all directly to the appropriate family
### Person File

<table>
<thead>
<tr>
<th>PERSON KEY</th>
<th>NAME</th>
<th>GENDER</th>
<th>DATE OF BIRTH *</th>
<th>FAMILY OF ORIGIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>P00337</td>
<td>ROBERT PATTISON</td>
<td>M</td>
<td>7 OCT 1819</td>
<td>-</td>
</tr>
<tr>
<td>P00338</td>
<td>FRANCES WILSON</td>
<td>F</td>
<td>7 OCT 1822</td>
<td>F00315</td>
</tr>
<tr>
<td>P00339</td>
<td>ANN PATTISON</td>
<td>F</td>
<td>7 OCT 1850</td>
<td>F00053</td>
</tr>
<tr>
<td>P00340</td>
<td>JOHN PATTISON</td>
<td>M</td>
<td>6 OCT 1852</td>
<td>F00053</td>
</tr>
<tr>
<td>P00341</td>
<td>FRANCES PATTISON</td>
<td>F</td>
<td>7 OCT 1853</td>
<td>F00053</td>
</tr>
<tr>
<td>P00342</td>
<td>ROBERT PATTISON</td>
<td>M</td>
<td>7 OCT 1855</td>
<td>F00053</td>
</tr>
<tr>
<td>P00343</td>
<td>EDWARD PATTISON</td>
<td>M</td>
<td>7 OCT 1857</td>
<td>F00053</td>
</tr>
<tr>
<td>P00344</td>
<td>MARY E PATTISON</td>
<td>F</td>
<td>7 OCT 1859</td>
<td>F00053</td>
</tr>
<tr>
<td>P00350</td>
<td>THOMAS N PATTISON</td>
<td>M</td>
<td>1 OCT 1861</td>
<td>F00053</td>
</tr>
<tr>
<td>P00351</td>
<td>EMMA PATTISON</td>
<td>F</td>
<td>1 OCT 1864</td>
<td>F00053</td>
</tr>
<tr>
<td>P01293</td>
<td>THOMAS WILSON</td>
<td>M</td>
<td>31 JUL 1778</td>
<td>-</td>
</tr>
<tr>
<td>P01294</td>
<td>ANN WILSON</td>
<td>F</td>
<td>3 JUN 1783</td>
<td>-</td>
</tr>
</tbody>
</table>

* calculated from the person's age, e.g. as specified in a census record.

### Family File

<table>
<thead>
<tr>
<th>FAMILY KEY</th>
<th>MARRIAGE PARTNER NAMES</th>
<th>DATE OF MARRIAGE</th>
<th>FATHER KEY</th>
<th>MOTHER KEY</th>
</tr>
</thead>
<tbody>
<tr>
<td>F00053</td>
<td>ROBERT &amp; FRANCES PATTISON</td>
<td>21 APR 1849</td>
<td>P00337</td>
<td>P00338</td>
</tr>
<tr>
<td>F00315</td>
<td>THOMAS &amp; ANN WILSON</td>
<td>-</td>
<td>P01293</td>
<td>P01294</td>
</tr>
</tbody>
</table>

Figure 11.6 The Population Database: a relational 'flat file' view
record, viz. for Robert and Frances Pattison, in the family file. This arrangement is equivalent to the 'P-CHILD-SET'. Consider next the family record for Robert and Frances Pattison. This contains two person key values, one for the father and the other for the mother. The value for the father is 'P00337', and this therefore links the family record directly to the person record for Robert Pattison. This arrangement is therefore equivalent to the 'P-FATHER-SET'. In a similar fashion the person key value for the mother provides a corresponding equivalent to the 'P-MOTHER-SET'. The inclusion of additional person and family records for Frances Wilson's family of origin (viz. 'F00315'), should serve to illustrate the simplicity with which a complex family structure may be collapsed into two flat files.

It is now possible to see precisely what needs to be done when off-loading person and family files. For each 'P-PERSON' record from the Population Database it is necessary to generate a person record containing the unique person key, together with other data items from the 'P-PERSON' record. Where the owning 'P-FAMILY' record in the 'P-CHILD-SET' is present in the database then it is necessary to retrieve this record so that its unique family identification key can be obtained and stored in the person record. The person record may then be written out to the person file.

When off-loading the family file it is necessary to locate each 'P-FAMILY' record in the Population Database and generate a corresponding family record. This will contain the unique family key, together with other data items from the 'P-FAMILY' record.
owning 'P-PERSON' record in the 'P-FATHER-SET' is present in the database then it is necessary to retrieve this record so that its unique person identification key can be obtained and stored in the family record. A similar operation must be carried out where the owning 'P-PERSON' record in the 'P-MOTHER-SET' is present. The family record may then be written out to the family file.

The off-loading of family structures from a CODASYL-type database in a form suitable for direct input to a relational database system is thus seen to be an eminently practicable proposition. If other records, such as 'P-EVENT' and 'P-CENS-EVENT', were provided with their own unique identification keys then the off-loading operation could be extended to encompass the whole of the Population Database. The opportunity to carry out a rigorous analysis of the Population Database would now be available.

I have completed my exploration of the problems of designing and implementing strategies for achieving the family-based linkage of nineteenth century demographic data. In the past five chapters I have examined in detail the nature of the functions which the major components of a linkage system, i.e. the subsystems, can contribute to the overall process of record linkage. And I have also examined the problems associated with the design of such subsystems and with the need for a progressive disaggregation of their functions, ultimately into a number of discrete, COBOL program modules.
In parallel with this functional analysis I have also observed the process of record linkage in terms of the progressive transformations of the nominal data. The effect of this has been to produce a partial shift in emphasis away from the linkage 'mechanisms' themselves and towards the overall systems 'architecture' in which the mechanisms operate. I shall explore this change in perspective more fully in Section 13.1.2.

The record linkage system whose development has provided a vehicle for the present research investigation has been used in a preliminary way to examine the demographic characteristics of a small study population. In the next chapter I shall assess the performance of the linkage system in its execution of this task.
NOTES

1. Examples of the display format were provided in Figures 6.3 and 6.4.

2. The source records which contributed to the displays referred to in Note 1 were illustrated in Figure 6.2.

3. A demonstration of this method of monitoring the linkage operations is provided in the next chapter.

4. It should be recalled that in population analysis displays all female names are given in the maiden name form. Where the maiden surname is not known it is replaced by the string '@@@@@@@@@', indicating missing information.

5. The 'true' maximum information would presumably be considered to consist of all the records in the Population Database which by whatever path are linked to the nominated 'P-PERSON' record. It would thus encompass every relative, every person in the same cohort, every person who has shared the same occupation or lived in the same enumeration district, etc., etc.

6. The wife in this display is, in fact, the Frances Wilson whose life-history was used for illustration in the previous section. It will be interesting, therefore, to observe the way in which the two sets of information overlap.
   It may also be noted that the family display serves to clear up a small mystery which may have been observed in the life-history display. In Figure 9.6 it was revealed that Robert and Frances Pattison had eight children, but also that Frances had been identified as mother in only seven baptisms. The information in Figure 11.2 makes clear why this was so. Thus, of the eight children in the family seven were born at Stotfold, which is within the parish of Elwick Hall, while the eldest, Ann Pattison, was born at Elwick, which is just outside the parish. Since the baptism records only for Elwick Hall parish were linked it is not surprising that Ann Pattison's baptism should not have been located.

7. As was mentioned in Chapter 9 a limited form of Place Index has also been provided, although the present implementation of the Population Analysis Subsystem does not support associated indexing commands.

8. An analysis of the 322 families in the database has revealed that only 11 of them can be connected to a 'genuine' 'P-FAMILY-COHORT' record, i.e. in only 11 cases are the marriage date and birthplace zone codes of both marriage partners known. Considering the selection criteria separately, for 74 families the marriage date is known and for 71 families the birthplace zone codes of both
marriage partners are known. These findings support the view expressed in Section 9.2.2 that some adjustment to the categories for family cohort selection may be called for.

9. The reason for this multiplicity of access routes and an account of the potential confusion which it can cause were presented in Chapter 9, Note 24.

10. The range of commands available was described in Section 6.6.

11. A description of a flat file was provided at the beginning of Section 4.2.1.

12. In a modified Population Database, redesigned for the enhanced version of the system, there would be no requirement for a Cohort Index, and so some simpler, alternative arrangement for accessing each 'P-PERSON' record could be provided.

13. While such unique keys are essential for the operation of relational systems they are provided in the current IDMS implementation only for presentational purposes and, in particular, to assist the user to cross-reference people and family displays.

14. Where it is appropriate, data items will still, of course, be held in an internal coded form.

15. Strictly, this name information need not be held (i.e. it is 'redundant'), since it can be obtained by accessing the person record of each marriage partner.

16. In an actual relational implementation each person and family key would consist of a simple integer value, i.e. there would be no prefix 'P' or 'F' character. The prefixes are retained in the current example for illustrative purposes, and, in particular, to facilitate comparison between the IDMS and relational schemes.
CHAPTER 12
THE STUDY POPULATION AND ITS ANALYSIS

The primary objective of the present research has been to explore the use of novel computing strategies to achieve the family-based record linkage of demographic data. This research has been carried out within the constraints imposed by a particular historical and geographical context, and in this chapter I shall begin by examining the characteristics of this context and the reasons for its selection. I shall then proceed to examine the 'nominal record universe' which has been employed as test data and the results of the linkages which have been performed. My main aim will be to establish the extent to which the linkage strategies which have been developed are capable of correctly linking nominal records. My secondary aim will be to make a preliminary substantive analysis of the linked data.

12.1 THE STUDY POPULATION

Within the original research objectives it was proposed, if time permitted, to carry out an associated piece of historical research. It was thought that such an exercise would serve both to subject the linkage strategies to a rigorous examination and validation and also to establish what kinds of results might be achieved. In the event, because the linkage issues were found to be so complex, there was
insufficient time available at the end of the project to exploit the system as fully as had been hoped. In this section I shall therefore begin by examining my original objectives and then look at the more modest study which was eventually carried out.

12.1.1 The Original Plan

The geographical area selected for analysis, which I have chosen to call the 'base zone' (1), consisted of the neighbouring rural parishes of Elwick Hall and Hart in County Durham. This choice was made in the first instance primarily on the basis of my familiarity with the area and the associated records (2) rather than for substantive reasons. However, a subsequent detailed analysis of the historical local context and an observation of the relevant parish register and other records confirmed that the setting was a suitable one in which to explore the methodological problems associated with nominal record linkage. In the first place, the parish registers for the base zone were available and complete, and there was no evidence of any lapse in the registration process in the proposed study period, viz. 1771-1871. Secondly, the total population of the base zone, viz. 957 in 1871 (3), was thought to be sufficiently large for a worthwhile study, and yet not so large that considerable resources would be needed for data preparation. Thirdly, there appeared to be little evidence of religious nonconformity (4), and attempts to link the baptism registers to the 1871 census gave moderately encouraging results (5). Finally, the geographical location of the zone presented
a potentially interesting way of investigating and partially overcoming the problem of migration, as I shall now demonstrate.

The positions of Elwick Hall and Hart parishes are illustrated in Map 12.1. While the population of the base zone rose only moderately during the nineteenth century, from 675 in 1801 (PP 1852-3 LXXXVI, 12) to 957 in 1871 (PP 1872 LXVI, 106), a dramatic increase occurred in the two coastal parishes immediately to the east, viz. Hartlepool and Stranton. In 1831 Hartlepool was a small fishing port with a population of 1330, and Stranton was a rural parish with a population of 736 (PP 1852-3 LXXXVI, 12). In 1832 the Hartlepool Dock and Railway Company was formed with the intention of providing facilities for the exporting of coal from the collieries ten miles to the north-west of Hartlepool (Wood 1967, 12-3). From this date onwards the two towns of Hartlepool and West Hartlepool grew rapidly, and by 1871 the total urban population was 38,252 (6). By this time the local industry was centred on the docks, shipbuilding and the production of iron and steel.

Since the development of the new urban area created job opportunities on a vast scale it was considered that the area would have provided a popular destination for outmigrants from the base zone. If this were true, and also if it were possible to locate records for the outmigrants, then it was thought that the adverse effects of population movement on record linkage might be minimised. If, by contrast, one were to select, for example, a study area which was remote from a particular centre of population then it would be
Map 12.1  The parishes of Elwick Hall and Hart and their environs
likely that the outmigration from such an area would be directed more
diffusely, and potentially towards many centres of population. Under
these circumstances the task of linking records for outmigrants would
be extremely severe.

In order to assess the extent to which outmigrants from the base
zone moved into the neighbouring urban area a search was made for them
in the census records for the neighbouring area. In a search of the
1871 census records for 12 enumeration districts in a section of
Stranton Parish with a population of 13,410 (7) the number of natives
of the base zone who were located was 56. If this number is scaled up
to provide an estimate of the number of natives of the base zone
present in the whole of the urban area then the value obtained is 160.
Since this is equivalent to 17% of the population of the base zone in
1851 (8) it represents a substantial movement of people into the urban
area. Moreover, one would obtain a significantly higher percentage
figure if one could locate in the urban area all the outmigrants from
the base zone, both native and non-native (9).

Given such results as these it was concluded that the selection
of a base zone which was immediately adjacent to a new urban area
could reduce the problems raised by population mobility. Thus, if a
large proportion of the rural to urban migration did take place into
the immediately adjacent urban area then potentially it becomes
possible to trace many of those who migrated. However, it cannot be
pretended that this 'solution' is especially elegant or simple,
requiring as it does extensive searching of records for the urban
area. Nevertheless, it does seem to offer a reasonably promising way to proceed, particularly in a situation where the historical context, involving large-scale population movements, presents serious obstacles to record linkage.

A final, important characteristic of the base zone which I shall now examine is its internal geographic and demographic composition. The population of the base zone in 1871 was 957, of which 217 lived in Elwick Hall parish and 740 lived in Hart parish. As can be seen in Map 12.1 Hart parish contains three villages: Hart, Elwick and Dalton Piercy. By contrast, Elwick Hall parish contains no villages or other significant centres of population, but only scattered farms and houses. Another important feature of Elwick Hall parish is the anomalous location of its church. As can be seen it is at the extreme north-east boundary of the parish and in very close proximity to the village of Elwick, in the neighbouring parish of Hart. The effect of this positioning is that during many periods of the church's history the inhabitants of Elwick and Dalton Piercy villages have used the Elwick Hall church in preference to their own church at Hart, which is two miles further away. Thus, for example, in the period 1861-70 only 37% of baptisms in the Elwick Hall register were for people resident in the parish; 54% were for people resident in Hart parish and 9% were for people resident elsewhere (DCRO 1978 Vol 2, 46-53). The position with regard to burials during the same period was even more marked. Only 21% of burials were for residents of Elwick Hall parish; 51% were for residents of Hart parish and 28% were for people resident elsewhere (DCRO 1978 Vol 2, 149-53). The practice of arranging for
burial in one's native or 'home' parish may partly account for this large proportion of non-resident burials. In contrast, the practice with regard to marriages was quite strict. In all cases in the period 1861-70 at least one of the marriage partners was stated to be resident in Elwick Hall parish (DCRO 1978 Vol 2, 100-2).

In view of the proximity of Elwick Hall parish church and Elwick village it was therefore regarded as important that the two parishes of Elwick Hall and Hart should be taken together as a single unit in the record linkage exercise if this was possible. The original plan did, however, envisage that the smaller of the two parishes, viz. Elwick Hall, would furnish an adequate amount of data for testing purposes during the system development. It was intended that at a later stage, and time permitting, data for Hart parish and selected data for the nearby urban area would also then be incorporated.

12.1.2 The Modified Plan

As was stated at the beginning of the chapter it became apparent that because of the complexity of the linkage issues it would be necessary to reduce the scale of the substantive element of the research. The same pattern seems to have occurred in the linkage work carried out by Macfarlane. As he observes: 'it is essential to stress that the whole task of community reconstruction is in essence extremely time consuming' (Macfarlane et al 1983, 29). And there appears to have been a consequent narrowing of focus during the
project from three communities to the single parish of Earle's Colne.

Within the present project the successful completion of the piece of research as originally proposed would have required the following:

1. the preparation of a significant amount of data and associated name directory records.

2. the development of improved population analysis facilities. Such a development, relying on the use of a relational database system, was outlined at the end of the last chapter. To embark on a major analysis of a population database without the provision of flexible retrieval facilities such as these could prove to be difficult and extremely time-consuming (10).

3. possibly some minor refinements to the record linkage strategies.

Clearly, significant resources would be required to accomplish these tasks. Indeed, the second task alone is a formidable one, and it could properly form the basis of a separate research investigation. Bearing all of these considerations in mind, therefore, it was decided that it would be prudent to work with a more modest set of substantive objectives. The characteristics of the modified plan which was subsequently adopted were as follows:
1. Data for the single parish of Elwick Hall was to be used (11). Thus, the proposal to add data for Hart parish and the urban area was withdrawn.

2. The study period was to be reduced from 1771-1871 to 1813-71. This served to reduce the task of data preparation, but also, and more significantly, it served to reduce the number of parish register data input formats which the linkage system was required to handle (12).

3. The census records to be linked were to include those for 1851, '61 and '71. The original research objectives also included a proposal to use the 1841 census data if time permitted (13).

It will be apparent that in substantive terms this modified plan is much less ambitious than the original one. Nevertheless, it was considered that the adoption of this plan would permit proper priority to be given to my major research objective, viz. the exploration of family-based record linkage strategies. At the same time, it was judged that the quantity and types of data included in the modified plan would be quite adequate for the important task of assessing the overall performance of the linkage strategies.
12.1.3 The Nominal Record Universe

The records for Elwick Hall parish which were entered into the system and linked are as follows:

   This total includes 18 illegitimate baptisms.
   N.B. For all parish register records the start date was taken as 1 January 1813 and the end date as 2 April 1871, i.e. the date on which the 1871 census was held.


4. 1851 census, 33 households, 187 persons (located in PRO H.O. 107/2384).

5. 1861 census, 37 households, 206 persons (located in PRO R.G./9 3699).

6. 1871 census, 34 households, 217 persons (located in PRO R.G./10 4913).

In addition, four selected households (24 persons) from the 1871 census for Stranton parish were included for testing purposes (see Note 18). In all the total number of records linked was 967.
12.2 VALIDATION OF THE LINKAGE STRATEGIES

A primary characteristic of the overall linkage strategy which has been employed is that the 'correctness' of the linkage should be measured retrospectively, i.e. by analysing the final contents of the Population Database (14). Two methods of doing this have been proposed. In the first it is necessary to take a number of records and link them manually. The results are then compared with those produced by the system. The second method involves some statistical analyses of the contents of the entire Population Database. Particular biases in the results may then, for example, indicate a tendency to 'over-link'. I shall now explore the use of these two methods in turn.

12.2.1 Manual Validation

Appendix C contains some sample linkage and analysis listings, and it is to these that I shall refer in the following discussion.

Figure C.1 contains a listing of all the nominal records present in the Source Database which contain a reference to the name 'ELDERS', or to a synonym of 'ELDERS'. Figures C.2 to C.4 illustrate the relevant contents of the Population Database after these records had been linked. By making a close analysis and comparison of these listings it should be possible to determine whether the linkage strategies have produced acceptable results (15).
I shall begin by examining the nominal records which are illustrated in Figure C.1. There are, in all, 25 records. These include 5 census household records, 5 marriages, 13 births and 2 deaths (16). It will be observed that the records are divided into two groups, the first containing records for the surname 'ELDERS', and the second containing records for its synonym 'ELDER'. Close examination of the records in each group will reveal that the two variants of the surname were used interchangeably for the same person. Thus, for example, the children William, Robert and Margaret S ELDERS in the first census household (C00012) are found to have the alternative surname spelling in their birth records (B00140, B00210 and B00231).

When an attempt was made to link the 25 records by hand the genealogical structures illustrated in Figure 12.1 were obtained. Unique person and family identifiers have been included in these illustrations in order to facilitate comparison with the structures generated by the system (17). And, as will now be demonstrated, when this comparison is carried out it is found that the structures generated by the system do closely match those produced by hand.

Let us examine Figures C.2 to C.4 in detail. Each figure presents a particular subset of the total information about the 'ELDERS' family. Figure C.2 provides a detailed display of each conjugal family. The first display, describing the family of John and Jane ELDERS (F00108), is a relatively simple one, created essentially from the information provided in just one marriage record (M00066).
Figure 12.1 Genealogical structures for the 'ELDERS' family, produced by manual linkage
However, the display additionally contains date and place of birth information extracted from census records. In the case of John ELDERS this information was obtained from an 1851 census household (C00001), shown in Figure C.1; in the case of Jane PORRIT the information was obtained from an 1861 census household (18).

The next conjugal family which is displayed in the listing, that of Thomas and Elizabeth ELDERS (F00023), is rather more elaborate. In this case there is no marriage information, but instead details of the nine children of the marriage. The information needed to construct the complete display was originally contained in three census households (C00012, C00068 and C00001), although in the case of John ELDERS (from household C00001) his connection to the family could be established only after his birth record (B00126) had also been encountered. The following conjugal family, that of Thomas and Hannah ELDERS (F00025), was created from information derived from just one census household (C00097).

All the families examined thus far correspond precisely with the manually created families illustrated in Figure 12.1. However, when we come to the final family in the display, that of William and Ann ELDER (F00205), we find that there is a minor discrepancy. In the system-produced version of the family Margaret ELDERS (P00557) has not been connected as one of the children. In the manually-produced version the connection has been made, although the dashed line is intended to indicate that the link is to be regarded as a tenuous one. The decision to make the connection was, in fact, based solely on the
information provided in Margaret ELDERS' marriage record (M00036), in which the name of her father is given as 'William'. And since, as described in Section 10.1.5, the present linkage strategy does not make use of the information identifying the father of each marriage partner, the reason for the missing link in the system-produced family structure is clear.

The next section of the population analysis listing, presented in Figure C.3, provides details about each of the people in the database called 'Thomas ELDERS'. In the original nominal records there are in all 21 references to 'Thomas ELDERS' (19), or to a synonym, and the function of this display is to confirm that the corresponding 21 events have been distributed 'correctly' to the appropriate individuals. Four people called 'Thomas ELDERS' have been identified (P00147, P00859, P00151 and P00160), and they match precisely the equivalent four people manually identified and illustrated in Figure 12.1. (Three of these people, it will be observed, are grandfather, father and son.) In addition, a close examination of the details for each individual confirms that all the events have been correctly distributed, as far as can be judged. In passing, it may also be noted that Thomas ELDERS (P00147) was successfully linked across eleven records, in three of which he was identified as 'Thomas ELDERS' and in the remaining eight of which he was identified as 'Thomas ELDER'.

The final section of the population analysis listing, presented in Figure C.4, consists of a nominal index of persons, and this
provides a convenient check-list against which the individuals in the manually-produced structures in Figure 12.1 can be cross-referenced. The index contains an alphabetically ordered list of all the people called 'ELDERS' or 'ELDER', and for each person there is displayed the unique person identifier, name, gender, and, where known, date and place of birth. Thus, for example, under 'THOMAS ELDERS' there are listed the four people whose details were more fully displayed in Figure C.3. For convenience in these displays the group of people sharing a particular name are listed chronologically in order of their dates of birth. Those for whom the date of birth is not known are placed at the end of the group. Under a female name are listed those for whom the name (or a synonym) is a maiden name, followed by those for whom the name (or a synonym) is a married name. Thus, for example, under 'ANN ELDERS' there are identified two people for whom this is the maiden name (P00856 and P00156) and two people for whom this is the married name (P00158 and P00858). Since for both of those in the second group the maiden surnames are not known these surnames are replaced by the special string '1@@@@@@@@'. In a situation where the maiden surname is known (e.g. Jane ELDERS, P00230) this is provided in the display.

A systematic comparison of the nominal index with the individuals presented in Figure 12.1 confirms that the manually-produced and system-produced structures are in precise agreement (20). My overall conclusion is therefore that, on the basis of this method of validation and from the results obtained in the particular example which has been examined, the linkage strategies are performing
satisfactorily.

Such a manual method of validation has also been used to assess the results of the person-based linkage of census and assessment data for the city of Hamilton, Ontario, Canada (Winchester 1970, 123). Agreement was established in only 95% of the cases, although this may be a reasonable result for an urban population and for the kinds of data and linkage strategies which were used.

12.2.2 Statistical Validation

The two main disadvantages of the manual method of validation are that it is an arduous, and necessarily error-prone, task and also that it is selective, and not comprehensive. A much more attractive approach would therefore be one which relied on the computer to carry out the validation and which subjected to observation the total contents of the Population Database (21).

It is not at present clear how a comprehensive scheme for computer validation might be organised. However, in this section I shall attempt, in an exploratory fashion, to employ a statistical technique which was proposed in Section 10.1.1 and which may exemplify the kind of approach which is envisaged. The technique to be investigated is specifically designed to detect the 'over-linkage' of census records, but it is anticipated that other statistical techniques could be devised to measure other aspects of the linkage.
The hypothesis which was proposed in Section 10.1.1 was that if there was a tendency for the linkage strategy to cause an over-linkage of census records, i.e. make connections which should not be made, then one would expect this to be particularly apparent for people with popular names, since for them there would be more candidate records available for linkage. It was also proposed that one should be able to detect the presence of over-linkage, in practice, by observing the discrepancies between the computed dates of birth derived from the census records for each individual: on average it would be expected that true links would display smaller computed date of birth discrepancies than false links. If, therefore, one were to find that people with popular names displayed significantly greater computed date of birth discrepancies than those with less popular names then it could be concluded that over-linkage had occurred. If, on the contrary, one were to find that the relative popularity of people's names had little or no bearing on the computed date of birth discrepancies then it could be concluded that over-linkage had not occurred. I shall now explore the use of this technique by examining the relevant contents of the Population Database (22).

In the whole of the Population Database 32 persons were identified in both the 1851 and 1861 censuses, 25 persons were identified in both the 1861 and 1871 censuses, and 14 persons were identified in all three censuses. The total number of inter-census linkages which may therefore be subjected to analysis is

\[32 + 25 + (2 \times 14) = 85.\]

Since this number of linkages is rather small it may be considered wise to subdivide it into only two groups
for the subsequent analysis. In the first group are included those
linkages for which the two computed dates of birth, when rounded to
the nearest year, are in agreement; in the second group are included
those for which there is a disagreement. When the 85 linkages are
subdivided in this way it is found that there are 59 (i.e. 69%) in the
first group and 26 (i.e. 31%) in the second group. These results are
illustrated in Table 12.1.

For the next step in the analysis it is necessary to determine
the relative popularity of the names of the people in the two groups.
Although this calculation should properly be based on the full name of
each person a more convenient and appropriate method of doing this,
especially for small population sizes, is to determine the relative
popularity of the individual surnames only. This may be done in the
present investigation by first drawing up a list of the surnames (23)
for the 71 persons for whom there are inter-census linkages: 22 such
names can be identified. The relative popularity of each surname is
then established by counting the number of occurrences of the name in
the households for all three censuses. Thus, for example, 'DARLING'
is found to be the most popular, with a count of 29, 'SMITH' the next
most popular, with a count of 28, and so on. The least popular are
'FIRBY' and 'LAIDLIER', each with a count of 5.

The final step in the analysis is to calculate a mean
'popularity' value for the two groups in Table 12.1. Thus, for the 59
linkages in the first group it is necessary to find the mean value of
the count of census occurrences for the 59 surname occurrences in this
<table>
<thead>
<tr>
<th>Number of linkages</th>
<th>Mean value of number of census occurrences for major surnames</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>No computed date of birth discrepancies</td>
<td>59</td>
<td>18.9</td>
</tr>
<tr>
<td>Computed date of birth discrepancies of one or more years</td>
<td>26</td>
<td>14.6</td>
</tr>
</tbody>
</table>

Table 12.1  The relationship between the presence or absence of computed date of birth discrepancies and the relative popularity of major surnames for census record linkages (1851-61 and 1861-71)

<table>
<thead>
<tr>
<th>Number of linkages</th>
<th>Mean value of age at first census</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>No computed date of birth discrepancies</td>
<td>59</td>
<td>27.1</td>
</tr>
<tr>
<td>Computed date of birth discrepancies of one or more years</td>
<td>26</td>
<td>39.4</td>
</tr>
</tbody>
</table>

Table 12.2  The relationship between the presence or absence of computed date of birth discrepancies and mean age for census record linkages (1851-61 and 1861-71)
group. The value obtained, viz. 18.9, together with the variance, are provided in the final two columns of Table 12.1. The values for the second group are calculated in the same way.

What then, if anything, can be concluded from the results illustrated in Table 12.1? Rather surprisingly, the group for which there is no computed date of birth discrepancy has a higher mean 'popularity' value than the other group. In other words, if one accepts the framework of the original hypothesis the conclusion appears to be that over-linkage is demonstrated, but that it is associated with the less popular surnames. A test for the significance of the result using the Student's t distribution (24) gives a t value of 2.55 and a confidence level of 1%, i.e. it is very unlikely that this result will have arisen by chance.

Since intuitively it appears absurd for over-linkage to be associated with the group for which there are fewer records (25) it seems clear that there must be some unexpected factor or factors present which are causing the results to be so strongly biased. A superficial observation of the data reveals that the ages of the people in each group could be the biasing factor. A further analysis of the two groups may therefore be carried out, but in this case the ages, rather than the surnames, of the members of each group are studied. For each link the age of the person at the earlier of the two censuses is noted and the mean obtained for the group. The results of this second analysis are illustrated in Table 12.2. It is quite evident that those in the group displaying no computed date of
birth discrepancies are considerably younger on average (approximately 12 years) than those in the other group. And the application of the Student's t test in this case gives a t value of 2.84 and an improved confidence level, viz. 0.5%.

The second set of results now provides a much more plausible explanation for the observed characteristics of the two groups. The predominant effect which the analyses appear to be measuring is the accuracy with which ages were reported, and the results seem to confirm that age reporting became less precise with increasing age. This is an intuitively reasonable result, since with increasing age, and therefore with increasing distance from the birth event, it seems probable that memory lapses and in some cases deliberate misreporting could have occurred. The only other observation which requires an explanation is why the group which is associated with the more popular names is also the group which contains younger people. The reason seems fairly self-evident. In a typical enumeration district the most common names will, in general, be those for which there are large family groupings. And large family groupings will normally consist at least of two young or middle-aged adults, together with a number of children. Clearly, the average age of such groups will normally be lower than for alternative family or household groupings.

It is clear that the attempt to detect the over-linkage of census records in the manner originally proposed has been undermined by an age bias in the data. In order to remove this bias it would be necessary to select for analysis two groups whose age characteristics
were comparable. At the same time it would also be advisable to introduce a more refined method of measuring the popularity of each name: ideally this method would be based on the use of full names, rather than surnames. However, given the modest size of the current Population Database it is not feasible to explore further this method of validation.

12.2.3 Observed Limitations in the Present Linkage Strategies

The task of carrying out the analyses described in this chapter has demanded the extensive observation of listings produced by the Population Analysis Subsystem. In pursuing this task a very small number of anomalies and potential linkage errors have been observed, and in this section I shall address these. In general, the overall logic of the linkage strategies has been found to be sound, and it is considered that by applying a small number of minor refinements the strategies are capable of a high level of reliability.

The only significant strategic refinement which is considered to be desirable concerns the method of matching persons and families. At present, when searching the Population Database for a particular person or family the search is terminated once a satisfactory instance is located. In practice this strategy has worked extremely well, particularly because of the informational richness of the data which was used, but also because the entities in the database which are interrogated during matching are not individual records but are
summations of earlier linkage operations, and so possess even greater informational richness. There were only two occasions where this strategy produced improbable links, both of which involved the linkage of parish register records.

A refinement to the present strategy would therefore involve the interrogation of all candidate person or family instances in a particular matching situation and the selection of the most likely one. However, it should be emphasized that the design of the kinds of match-scoring algorithms which this would entail could involve considerable complexity, since for each person or family instance in the Population Database the associated information which is available can be in a virtually limitless range of states. By contrast, match-scoring algorithms used for person-based record linkage are normally required only to make comparisons of a handful of parallel fields in pairs of records (Winchester 1970, 117-22).

Apart from the inclusion of the above refinement, other modifications are considered necessary merely to correct trivial programming errors or to incorporate additional, simple checking mechanisms. For example, the linkage error identified during the manual validation exercise (26) was caused because of a missing consistency check. The incorporation of such checks are, however, a trivial programming task and have little bearing on the overall design of the linkage strategies.
12.3 A PRELIMINARY SUBSTANTIVE ANALYSIS OF THE LINKED DATA

In the last section it was observed how my revised research plan enabled the overall performance of the linkage system to be assessed. In this section I shall proceed to establish whether this plan is also able to furnish any significant substantive results.

The task of deriving substantive results from a record linkage project is not a trivial exercise. Even the task of off-loading the results of the linkage, a necessary prerequisite of any analysis, requires substantial effort. Thus, as Macfarlane has observed: 'The gathering together and indexing of the records using both hand and computerized methods has been a much larger task than we originally anticipated.' (Macfarlane et al 1983, 37).

Some introductory comments should first be made about the way in which the analysis has been carried out. In the absence of more sophisticated analytical tools, such as might be provided by a relational database interface, it has been necessary to carry out the analysis by sifting and searching through about 500 pages of listings produced by the Population Analysis Subsystem. The major part of these listings consists of a detailed display of information about every person in the Population Database. This display is subdivided into five sections, containing the following groups of people:
The study population and its analysis

- male natives of the base zone
- female natives of the base zone
- male non-natives of the base zone
- female non-natives of the base zone
- all who cannot be classified in the previous four sections

For each of the first four sections persons are displayed chronologically in order of their dates of birth. The start of the display for the female natives of the base zone is illustrated in Figure 12.2. Two other listings have also been used. The first is a listing of all the families in the database (27). The second is an index of occupations; a section of this index is illustrated in Figure 12.3. From this index one would normally go to the more detailed person displays to obtain the required family and life-history information.

In the analyses which follow I begin, in a fairly general way, by examining the degree to which inter-record linkage has actually taken place. For the sizeable group of people who can be identified in only one record an attempt is made to discover why no record linkage has taken place for them. I then turn my attention to those who have been identified in two or more records, and initially observe the duration
### Ch. 12

**The study population and its analysis**

<table>
<thead>
<tr>
<th>Year</th>
<th>Name</th>
<th>Gender</th>
<th>Date</th>
<th>Location</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1792</td>
<td>Elizabeth</td>
<td>F</td>
<td>27 Sep 1792</td>
<td>Wulviston</td>
<td>(P)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Father</td>
<td>NOT KNOWN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mother</td>
<td>NOT KNOWN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HUSBAND</td>
<td>P00283</td>
<td>GEO NEWBY</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SONS</td>
<td>1</td>
<td>DAUGHTERS</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1851 CENSUS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAR 1851</td>
<td>WIFE</td>
<td>MAR</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C00020</td>
<td>20</td>
<td>ELWICK HALL PARISH</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1861 CENSUS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>APR 1861</td>
<td>MOTHER</td>
<td>WID</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>C00057</td>
<td>24</td>
<td>ELWICK HALL PARISH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1794</td>
<td>Jane</td>
<td>F</td>
<td>28 Sep 1794</td>
<td>Elwick Hall</td>
<td>(B)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Father</td>
<td>NOT KNOWN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mother</td>
<td>NOT KNOWN</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 12.2** The start of the display of person details for all female natives of the base zone by the Population Analysis Subsystem.
### The study population and its analysis

#### Ch. 12

<table>
<thead>
<tr>
<th>CARTER</th>
<th>ALL OCCURRENCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>P00529</td>
<td>HENRY CORNER M 7 OCT 1833 ← 183D COOPEN BEWLEY (N)</td>
</tr>
<tr>
<td>P00055</td>
<td>WILLIAM FENWICK M 7 OCT 1842 ← 183D WHORLTON (N)</td>
</tr>
<tr>
<td>P00368</td>
<td>JOHN KITCHEN M 7 OCT 1843 ← 183D YARM (N)</td>
</tr>
<tr>
<td>P00056</td>
<td>MICHAEL JACKSON M 6 OCT 1848 ← 183D WOLVISTON (P)</td>
</tr>
</tbody>
</table>

#### PP/SUBSET=PERSONS/JOB=CARTWRIGHT

<table>
<thead>
<tr>
<th>CARTWRIGHT</th>
<th>ALL OCCURRENCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>P00885</td>
<td>STEPHEN GIBBON M</td>
</tr>
<tr>
<td>P00169</td>
<td>GEORGE FIBBY M 6 OCT 1820 ← 183D HIFSWELL (N)</td>
</tr>
<tr>
<td>P00183</td>
<td>BINKS FURBY M 29 SEP 1834 ← 183D HIFSWEL (N)</td>
</tr>
</tbody>
</table>

**Figure 12.3** A section of the index of occupations produced by the Population Analysis Subsystem
for which such individuals were 'in view'. Finally, I briefly examine some occupational and other demographic characteristics of the people and families in the Population Database.

12.3.1 Event Analysis

Each record from the Source Database contributes one or more 'events' to the Population Database. In the process of record linkage these events become associated with other events, and in this way they participate in the formation of life-histories for the individuals in the Population Database. An interesting way therefore of viewing the record linkage process is to observe the group-forming properties of the individual events.

The universe of records which was entered into the system and linked was defined in Section 12.1.3. From this definition it is a simple matter to calculate the total number of events which are created in the Population Database when these records are linked. For a burial record and for each person in a census household one event is created. For a marriage or an illegitimate baptism two events are created, while for a legitimate baptism three events are created. Applying these multiplication factors to the 967 records defined in Section 12.1.3 it is found that the corresponding number of events is 2407.
For an elementary analysis of these 2407 events one can examine the way in which they are distributed to the 1563 persons in the Population Database. The results of this analysis are illustrated in Table 12.3. The first row of this table reveals, rather surprisingly, that 74% of the persons in the Population Database are identified in only one record and that 48% of all events are not linked to other events. In addition, it reveals that only 4% of the population are identified in five or more records. In the next two sections I shall examine what lies behind these gross figures, and attempt to account for the apparently low linkage statistics.

12.3.2 The Unlinkable Persons

Since the group of persons for whom only one event has been identified constitute the majority of the population it will be instructive to analyse the characteristics of this group in more detail. Of the total 2407 events created in the Population Database 1157 were not linked to any other event. For this group of 1157 events Table 12.4 illustrates for each type of event the number and percentage which were not linked. Thus, for example, baptism events are shown to have the lowest capacity for linkage: from the total of 429 baptism events 305 (i.e. 71.1%) were not linked to any other events.

Let us first examine in detail the 'unlinkable' people who were identified in census records. From the total of 354 as many as 336
The study population and its analysis

<table>
<thead>
<tr>
<th>Number of events linked per person</th>
<th>Persons</th>
<th></th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>1157</td>
<td>74.0</td>
<td>1157</td>
</tr>
<tr>
<td>2</td>
<td>230</td>
<td>14.7</td>
<td>460</td>
</tr>
<tr>
<td>3</td>
<td>69</td>
<td>4.4</td>
<td>207</td>
</tr>
<tr>
<td>4</td>
<td>44</td>
<td>2.8</td>
<td>176</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>1.5</td>
<td>115</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>0.7</td>
<td>66</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>1.1</td>
<td>119</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
<td>0.4</td>
<td>56</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.1</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.1</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>0.1</td>
<td>22</td>
</tr>
<tr>
<td>All</td>
<td>1563</td>
<td>99.9</td>
<td>2407</td>
</tr>
</tbody>
</table>

Table 12.3 The distribution of events to persons in the Population Database
<table>
<thead>
<tr>
<th>Type of event</th>
<th>Total number of events</th>
<th>Events not linked</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1851 census</td>
<td>187</td>
<td>94</td>
<td>50.3</td>
<td></td>
</tr>
<tr>
<td>1861 census</td>
<td>206</td>
<td>96</td>
<td>46.6</td>
<td></td>
</tr>
<tr>
<td>1871 census</td>
<td>241</td>
<td>164</td>
<td>68.0</td>
<td></td>
</tr>
<tr>
<td>Baptism</td>
<td>429</td>
<td>305</td>
<td>71.1</td>
<td></td>
</tr>
<tr>
<td>Father in baptism</td>
<td>411</td>
<td>80</td>
<td>19.5</td>
<td></td>
</tr>
<tr>
<td>Mother in baptism</td>
<td>429</td>
<td>88</td>
<td>20.5</td>
<td></td>
</tr>
<tr>
<td>Marriage</td>
<td>148</td>
<td>80</td>
<td>54.1</td>
<td></td>
</tr>
<tr>
<td>Burial</td>
<td>356</td>
<td>250</td>
<td>70.2</td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>2407</td>
<td>1157</td>
<td>48.1</td>
<td></td>
</tr>
</tbody>
</table>

Table 12.4 The proportion of events which were not linked for each event type
(i.e. 95%) were recorded as being not born in Elwick Hall parish. Clearly, for them it is most unlikely that there will be linkage to their baptism records. Of the 260 un-linkable persons identified in the 1861 and 1871 censuses 62 (i.e. 24%) were born in the 1851-61 inter-census years and 44 (i.e. 17%) were born in the 1861-71 inter-census years. For the majority of this 41% group who were born after the 1851 census the scope for linkage to other records would be minimal. In particular, those born in the 1861-71 period (and, because of their presence in the census, guaranteed to be still alive in 1871) could potentially have been linked only to a baptism record.

Let us next examine the baptism events for which there were no linkages to other events. Of these 305 events 180 (i.e. 59%) were for people who were not resident in Elwick Hall parish. For the reasons given in Section 12.1.1 the probability of being able to link these events to any others, apart perhaps from burials, must be considered low.

The 'father in baptism' and 'mother in baptism' events appear to have a much better capacity for linkage than any others. Thus, only 20% of these events were not linked. The explanation for this result is fairly straightforward. Since for a typical family, with two or more children, the interval between births is likely to be between one and three years, there is a good probability that most families will have remained resident in one place long enough (e.g. two years) to have had at least two baptisms entered in the parish register. By contrast, in order for a family to be linked from one census to the
next it would need to have remained resident in one place for between 10 and 20 years.

Since marriage ceremonies at the Elwick Hall parish church were strictly permitted only for residents of the parish, the corresponding marriage records would appear to be the only parish records which are unlikely to be adversely affected by the anomalous location of the church. If this should be the case then for those referred to in the marriage records one would expect to be able to find links to both the other parish register records (viz. baptisms and burials) and the census records. These expectations appear to be borne out by the relatively good linkage results for marriage events in Table 12.4: of 148 events only 80 (i.e. 54.1%) were not linkable. Of the 80 which were not linkable 26 (i.e. 33%) were for non-residents of the parish and so would be unlikely to link to earlier baptism or census records. Since also 23 of this group of 26 events were for male non-residents it would seem likely that most of these would have subsequently taken their brides away from the parish and so would be unlikely to have linkages to other records.

Let us finally examine the burial events. Their capacity for linkage appears to be almost as low as for the baptism events: thus, 70.2% were not linked to any other events. Of the 250 which were not linkable 159 (i.e. 64%) were for persons born before 1813, and so for these there would be no possibility of linkage to baptism records. 163 (i.e. 65%) of the events were for persons buried before 30th March 1851, and for these there would be no possibility of linkage to census
records. Finally, of the 87 persons who were buried after 30th March 1851 77 had residences which were not within Elwick Hall parish. Thus, of the 250 unlinkable burial events one would confidently anticipate being able to link only 10 of them (i.e. 4%) to a census event.

These analyses illustrate clearly that there are several contributors to the apparently low linkage statistics. Population mobility is an intrinsic problem in any record linkage operation, and this has obviously been a major factor. The anomalous location of the parish church, together with the practice of providing baptisms and burials for non-residents of the parish, has exacerbated the situation. A final, intrinsic contributor to the low linkage statistics is that many people reported earlier in the parish registers did not survive long enough to appear in the census returns, and similarly those who were born towards the end of the study period were too young to appear in more than one or two records.

12.3.3 The Linkable Persons

In this section I shall make an initial examination of the 26% of the population (i.e. 406 persons) who were identified in more than one record.

From a methodological viewpoint it is important to establish the duration for which the individuals who were linked were 'in view':
whether or not a significant longitudinal study of a population will be possible is crucially dependent on this. For present purposes the 'in view' duration for each relevant individual can be obtained by finding the difference between the dates of the person's chronologically first and last events (28). The results so obtained are illustrated in Table 12.5, and from these it is evident that the vast majority of the 'linkable' section of the population were in view for a relatively short period. Thus, almost 60% were in view for less than 10 years, and only 11% were in view for more than 24 years. When these figures are related to the whole population one finds that only 2.9% were in view for more than 24 years. Clearly, with such results as these one must conclude that the scope for mounting a significant longitudinal study of such a population is remote (29).

Finally, there is another group which has an intrinsic interest from a methodological viewpoint, and this is the group of persons for whom their entire life-span is in view. For each such person in the present context the implication is that both the baptism and burial events have been successfully linked. This property exists for 44 of the linkable persons (i.e. 11%), most of whom are found to be children who died in infancy or childhood. The mean life-span of this group is 5.0 years and only six have a life-span of more than 12 years (30).

In the next two sections I shall briefly examine a number of the demographic characteristics of the Population Database. I shall first focus on occupational aspects and then look at the characteristics associated with the family.
In view
duration in
years*

<table>
<thead>
<tr>
<th>'In view' duration in years*</th>
<th>Persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>163</td>
</tr>
<tr>
<td>5-9</td>
<td>76</td>
</tr>
<tr>
<td>10-14</td>
<td>67</td>
</tr>
<tr>
<td>15-19</td>
<td>34</td>
</tr>
<tr>
<td>20-24</td>
<td>21</td>
</tr>
<tr>
<td>25-29</td>
<td>16</td>
</tr>
<tr>
<td>30-34</td>
<td>14</td>
</tr>
<tr>
<td>35-39</td>
<td>4</td>
</tr>
<tr>
<td>40-44</td>
<td>4</td>
</tr>
<tr>
<td>45-49</td>
<td>4</td>
</tr>
<tr>
<td>50-54</td>
<td>3</td>
</tr>
<tr>
<td>All</td>
<td>406</td>
</tr>
</tbody>
</table>

* N.B. Since the study period was taken from 1st January 1813 to 2nd April 1871 the maximum possible 'in view' duration is 58 years.

Table 12.5 The 'in view' duration for persons who were identified in more than one record.
12.3.4 An Analysis of Occupations

Since according to Table 12.5 there are 167 persons in the Population Database who are in view for more than nine years this would indicate that there is scope for studying individual occupational mobility. But, as was discussed in Section 3.2.1, there are problems associated with the use of occupational descriptions, and so these must first be addressed.

An interesting illustration of one of the problems of studying occupational mobility is provided in the life-history display of Thomas ELDERS (P00151) in Figure C.3. Between the censuses of 1861 and 1871 there are five descriptions of his occupation, and they are as follows:

1861 CENSUS: AG LAB
1866 FATHER/BAP: LABOURER
1868 FATHER/BAP: WOODMAN
1870 FATHER/BAP: WOODMAN
1871 CENSUS: FORESTER

A superficial observation of this display might lead one to conclude that Thomas ELDERS had undergone three changes in his occupation during the 1861-71 period. What seems more probable, however, is that there was only one change: from 'agricultural labourer' to 'woodman'.
Firstly, let us consider the apparent change from 'agricultural labourer' at the time of the 1861 census to 'labourer', as was entered in the baptism register in 1866. This difference can be accounted for quite simply by the fact that the methods of reporting occupations in the census and the parish registers are not identical (31). The occupational title 'agricultural labourer' is not used in the parish registers; instead the more imprecise title 'labourer' appears to be generally used. In the census, by contrast, the occupational title 'labourer' is used only rarely, and 'agricultural labourer' appears to be the title which is normally used. The apparent transition of Thomas ELDERS' occupation from 'woodman' to 'forester' can be explained by a similar mismatch in the methods of reporting. In particular, it is noteworthy that, while in the census both the titles 'woodman' and 'forester' are used, the title 'forester' is not used in the parish registers. Another significant discrepancy in the two methods of reporting concerns the precision with which domestic servants are described. In the census a whole range of discrepancies is used: 'domestic servant', 'housemaid', 'kitchen maid', 'cook', and so on. In the parish registers the generalised term 'servant' appears to be the one commonly used to cover all such occupations.

It is evident that any attempt to study occupational mobility which makes use of disparate source materials must be approached with particular caution. Therefore, in view of the above-mentioned problem and of the rather restricted data available it was decided that a thorough-going analysis based on the occupational descriptions was unlikely to be particularly fruitful. In the remainder of this
section, therefore, I shall merely make a few brief comments on a number of observations which have been made and which relate to the occupational descriptions.

A simple method of studying occupational mobility, and one which largely avoids the problems of using disparate sources, is to examine only the inter-census linkages (32). Unfortunately, in the present Population Database there are only 30 such linkages for working males, and so this is too small a sample from which to derive any significant results. But, in passing, it may be mentioned that for this group, each of whose members clearly enjoyed a strong residential continuity (i.e. by staying in one place for at least 10 years), there is a strong measure of occupational continuity. 20 of the 30 linkages display no change in occupational title, and of the remaining 10 linkages 5 display either a potentially cosmetic change in title (e.g. from 'forester' to 'woodman') or an internal upgrading (e.g. from 'groom' to 'coachman'). These observations are consistent with some analyses of occupational mobility, based on the manual family reconstitution of the parish registers (1700-1840) of the very large rural parish of Hawkshead in Lancashire (Oosterveen 1974). This study, using 20% of the parish records, revealed that 83% of male heads of families (i.e. 300 out of 363) consistently registered the same occupation.

By examining the occupations of fathers and sons in the Population Database it is possible to measure the amount of inter-generational occupational mobility which took place. Data on 30
such father-son relationships is available, and this is presented in Table 12.6. For simplicity occupations are grouped into four major categories: professional, farmer, trade and labourer, the labourer category subsuming all unskilled occupations, the three main subcomponents being labourer/agricultural labourer, hind and woodman/forester. For 20 of the 30 relationships the occupational category for father and son is identical, indicating a low rate of inter-generational occupational mobility (33). It is also noteworthy that only two sons of the professionals and farmers are in the trade and labourer categories, and that only two fathers in the trade and labourer categories have sons who are not also in the trade and labourer categories.

Finally, it is possible to examine the relationship between occupation and age at death. Although in view of the limited amount of data available only an elementary analysis is possible. Table 12.7 presents data on ages at death for the two major occupational categories: farmer and labourer (34). The results indicate that the farmers lived on average about eight years longer than the labourers. However, a test for the significance of this conclusion using the Student's t test gives a t value of 1.07 and a confidence level of 30%, i.e. it is quite possible that this result will have arisen by chance. But even if the result is significant, some caution is needed in its interpretation. Thus, if, as seems likely, there was some occupational mobility from 'labourer' to 'farmer' then the probable result of this would be that the mean age of the farmers in the population would be greater than the mean age of the labourers. Under
### Table 12.6 Inter-generational occupational mobility: the relationship between the occupations of fathers and their sons

<table>
<thead>
<tr>
<th>Occupation of son.</th>
<th>Occupation of father</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional</td>
<td></td>
<td>2</td>
<td>67</td>
</tr>
<tr>
<td>Farmer</td>
<td></td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Trade</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Labourer</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>All</td>
<td></td>
<td>3</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Occupation of father</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>Farmer</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>Trade</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>Labourer</td>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td>All</td>
<td>30</td>
<td>99</td>
</tr>
</tbody>
</table>

### Table 12.7 Age at death for farmers and labourers

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Number of persons</th>
<th>Mean age at death</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farmers</td>
<td>12</td>
<td>56.3</td>
<td>276</td>
</tr>
<tr>
<td>Labourers</td>
<td>9</td>
<td>48.0</td>
<td>296</td>
</tr>
</tbody>
</table>

Table 12.6: Inter-generational occupational mobility: the relationship between the occupations of fathers and their sons.
these circumstances it would not be surprising if one were to detect an apparent longevity in the farmers.

12.3.5 An Analysis of Family Characteristics

In the first part of this section I shall examine the available data on age at first marriage. 18 families in the Population Database are eligible for this analysis, i.e. they are families for which the marriage record is available, age information is available and for which both partners were single prior to the marriage. The results of the analysis are presented in Table 12.8. The conclusion that grooms were on average 4.5 years older than their brides is found to be significant at the 5% level, suggesting that it is unlikely that this result will have arisen by chance (35).

When the above 18 families are subdivided according to the occupation of the groom it is found that six of the grooms are farmers and five are labourers. The results of an analysis of the ages at first marriage for these two groups are presented in Table 12.9. The farmers are shown to be considerably older than the labourers at their first marriages (viz. about 12 years), and this conclusion is found to be significant at the 2.5% level. Again, this suggests that it is unlikely that such a result will have arisen by chance. However, as for the comparable results concerning age at death presented in the last section, the interpretation of these results also requires caution. Thus, once again the observed discrepancy in age at first
### Table 12.8 Age at first marriage for grooms and brides

<table>
<thead>
<tr>
<th></th>
<th>Number of persons</th>
<th>Mean age at first marriage</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grooms</td>
<td>18</td>
<td>30.5</td>
<td>78.0</td>
</tr>
<tr>
<td>Brides</td>
<td>18</td>
<td>26.0</td>
<td>33.9</td>
</tr>
</tbody>
</table>

### Table 12.9 Age at first marriage for farmers and labourers

<table>
<thead>
<tr>
<th></th>
<th>Number of persons</th>
<th>Mean age at first marriage</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farmers</td>
<td>6</td>
<td>36.2</td>
<td>92.5</td>
</tr>
<tr>
<td>Labourers</td>
<td>5</td>
<td>23.8</td>
<td>7.4</td>
</tr>
</tbody>
</table>

### Table 12.10 Geographical mobility for the families of farmers and labourers

<table>
<thead>
<tr>
<th></th>
<th>Number of families</th>
<th>Mean number of residences</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farmers</td>
<td>14</td>
<td>1.71</td>
<td>0.92</td>
</tr>
<tr>
<td>Labourers</td>
<td>9</td>
<td>3.22</td>
<td>1.06</td>
</tr>
</tbody>
</table>
marriage could be directly attributable to some unidirectional mobility between the two occupational categories.

For 17 families it is possible to calculate the interval between the date of marriage and the computed date of birth of the first child of the marriage. This interval ranges from -2 months to 38 months (36), with a mean value of 13.4 months and a median value of 13 months.

The geographical mobility of a family can be conveniently studied by analysing the birthplace information for each child of the family. This has been done, for example, by examining 1851 census data (Anderson 1971, 40-1). In order to compare the relative mobilities of the families of farmers with those of labourers it was decided to select for analysis those families in the Population Database for which there was a moderately large number of children (in fact, five or more) and determine the number of distinct residences which occurred for each of these families by analysing the birthplace information for the children. The results are presented in Table 12.10. There is a clear indication that on average each labourer family was almost twice as mobile as each farmer family, when measured in terms of the number of residences occupied. And this result is found to be significant at the 0.5% confidence level, i.e. it is very unlikely that it will have arisen by chance (37). Despite the observed high mobility for labourer families it is evident from the birthplace information that the distances travelled were usually small, involving typically the movement from one farm to another,
perhaps little more than one or two miles away.

12.4 SYSTEM PERFORMANCE

An important property of any system, and one which has so far been given little consideration, is its operational efficiency. In this final section of the present chapter I shall therefore briefly examine the computing resources which were needed to create and display the Population Database whose characteristics have been analysed above.

In Table 12.11 are displayed the processor times which were required to carry out the various stages of the total record linkage operation. The time required for the whole operation was thirteen minutes, of which 36% was actually involved in linking records. The 29% of the time spent in population analysis was used to produce the comprehensive displays which were employed in the analyses described in Section 12.3.

An interesting statistic is the time required to link a single record. The total number of records linked was 967 and the time needed to translate, load and link these records was 435 sec. The time required to link a single record is therefore 0.45 sec., which seems moderately efficient for an application of this complexity on the type of machine used (viz. ICL 2980).
<table>
<thead>
<tr>
<th>Operation</th>
<th>Processor time required in seconds</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database initialisation</td>
<td>65</td>
<td>8</td>
</tr>
<tr>
<td>Directory prime</td>
<td>55</td>
<td>7</td>
</tr>
<tr>
<td>Source translation</td>
<td>117</td>
<td>15</td>
</tr>
<tr>
<td>Source load</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>Record linkage</td>
<td>282</td>
<td>36</td>
</tr>
<tr>
<td>Population analysis</td>
<td>225</td>
<td>29</td>
</tr>
<tr>
<td>All</td>
<td>780</td>
<td>100</td>
</tr>
</tbody>
</table>

* N.B. All operations were performed on the Edinburgh Regional Computing Centre's ICL 2980 machine running under the VME/B operating system.

Table 12.11 The processor time required to carry out each stage of the total record linkage operation
Finally, one should take note of the amount of storage which was used by the IDMS database system to hold the complete database. This was almost 1.5 Mbytes, of which about 1 Mbyte was needed for the Population Database, and the remaining 0.5 Mbytes was split approximately equally between the Directory and Source Databases (38). Although IDMS employed no record compression techniques its method of storage allocation was economical.

I shall now draw together my main conclusions from the observations and analyses described in this chapter. They are as follows:

1. The overall linkage strategies have been shown to perform satisfactorily when presented with relatively small quantities of data. A small number of minor limitations have been observed in the strategies, but these could be overcome by making appropriate, mostly trivial, modifications.

2. A preliminary attempt to validate the results of the linkage by means of a statistical analysis of the contents of the Population Database was unsuccessful. This attempt was undermined by an age bias in the data used and by an insufficiency of data, generally.
3. The observation that almost three-quarters of the people in the Population Database could be identified in only one record confirms the difficulties of mounting such substantive studies of nineteenth century populations. In view of certain characteristics of the particular study population which was used (e.g. its limited size) this conclusion may, however, be over-pessimistic.

4. An additional, somewhat underestimated problem concerns the potential unreliability of using for analysis linked data originating from disparate source materials. Thus, for example, it was found that the adoption of different methods of reporting occupations in the census and the parish registers could cause spurious conclusions to be reached about occupational mobility.

5. A number of tentative, substantive conclusions were reached, of which the following appear to be the most reliable. Firstly, there was good evidence that age reporting in the census became less precise with increasing age. Secondly, there appeared to be a strong, positive correlation between the occupations of fathers and the occupations of their sons, i.e. little evidence of inter-generational occupational mobility. Thirdly, there was an observed mean age difference of 4.5 years in the age at first marriage for grooms and brides: the mean ages were 30.5 for grooms and 26.0 for brides. Finally, there was a clear indication that on average
each labourer family was almost twice as geographically mobile as each farmer family. At the same time, it was discovered that the distances travelled by the labourer families were usually small.

In the next and final chapter I shall return to the central, methodological issues which have been my main concern in this research investigation, and shall make my overall assessments and conclusions.
NOTES

1. The concept of the base zone as 'the immediate study area in which one is interested' was introduced in Section 3.2.2.

2. During the period 1969-78 I transcribed, edited and indexed the parish registers of Elwick Hall parish (DCRO 1978).

3. PP 1872 LXVI, 106. The Parish of Hartlepool (population 13,166), although listed in the tabulation for Hart parish, was not included in the base zone. Similarly, Throston township (population 1918), substantially a small urban concentration adjoining Hartlepool, was excluded, apart from 12 scattered rural households (78 individuals).

4. The parish registers themselves display little evidence of nonconformity, apart from the occasional reference to the baptism of Quakers. In the Elwick Hall baptism register (DCRO 1978 Vol 2, 12) there was one such baptism in 1783, and in 1820 six Quakers, all children of one family, were baptized (ibid, 25). Additionally, there is evidence elsewhere of a Wesleyan Methodist presence towards the end of the study period. Thus, in 1867 a small Wesleyan chapel, with seating for about 70, was built in Elwick village in the parish of Hart. And in the 1871 census (PRO R.G./10 4913, Township of Hart, p 8) Jonathan Moody, aged 35, a native of Hart, has his occupation described as 'Railway Station Master & Wesleyan Local Preacher'. It seems probable that he would have officiated at the recently built chapel at Elwick, two miles away.

5. In the 1871 census for Elwick Hall parish and Elwick township (PRO R.G./10 4913) 118 people were recorded as having been born in one or other of the two parishes. A search of the baptism registers of Elwick Hall parish (DCRO 1978 Vol 2, 6-53) and Hart parish (DCRO 1676, 91-189; DCRO 1813, 1-99; Hart Parish 1857, 1-29) revealed corresponding baptism entries in 90 instances, i.e. the success rate was 76%. (This calculation excludes married women (13 instances), for whom linking to baptism entries would have been more difficult.)

For comment on a similar attempt, by Wrigley to link baptism register entries to census data see Chapter 3, Note 19.

6. PP 1872 LXVI, 106-7. The total urban population consisted of 23,246 in Stranton parish, 13,166 in Hartlepool and 1840 in the urban section of Throston township.

7. Enumeration districts located in PRO R.G./10 4918, 4919 and 4921.

8. PP 1852-3 LXXXVI, 12. The population of the base zone in 1851 was 944. As in the corresponding calculation for 1871 (see Note 3) Throston township was excluded, apart from those people living in
scattered rural households: in this case there were 12 such households (77 individuals).

9. Unfortunately, the location of the non-native migrants is difficult, except where they occur in a household which contains at least one native migrant. Thus, for example, the location of a family with an outmigrant native child implicitly also locates outmigrant parents, since it can reasonably be assumed that, in general, the parents would also have been living in the base zone at the time of the child's birth.

10. To complete even the modest analyses described later in this chapter took many hours and required much sifting and searching through two inches (more than 500 pages) of listings produced by the Population Analysis Subsystem.

11. Four selected households from the 1871 census for Stranton parish (obtained from PRO R.G./10 4918, 4919 and 4921) were also included for testing purposes. The final event illustrated in the life-history of 'FRANCES WILSON' in Figure 9.6 derives from one of these households, and its presence serves to indicate that she was a migrant from Elwick Hall parish to the urban area.

12. The system as currently implemented can handle one input format for baptisms, two input formats for marriages and one input format for burials. Examples of these formats are provided in Appendix B. Were it required to extend the start of the study period back to 1771 then enhancements would need to be made to the system in order to accommodate three new record input formats. A new format would be required for baptism records in the period 1798-1812, because of the highly detailed form of the entries during this time. (An example is provided in Section 3.3.1.) For the same reason a new format for burial records in the period 1798-1812 would also be needed. Finally, the content of burial records in the period 1771-97 would demand an additional format. During this period there was normally no age at death provided (unlike post-1797 records), but often there was additional information, such as marital status and/or name of surviving spouse, which was not included in post-1812 records.

13. Again, the chief difficulty with incorporating the 1841 census data would not be the additional data preparation involved, but rather the requirement to enhance the system to handle a new data format. The same input format and linkage strategy can be used for each of the post-1841 censuses, because each has the same information content and structure. However, the 1841 census data has a sufficiently different content and structure (e.g. lacking relationship and detailed birthplace information) to require its own input format and linkage strategy.


15. This particular family and its associated records was selected for demonstration purposes for a number of reasons. In the first
place the family name (viz. 'ELDERS') is a comparatively common one, and so there is a moderately large number of records to be linked. Secondly, the records to be linked include examples taken from the whole range of data input formats. Thirdly, the family name has a commonly used synonym (viz. 'ELDER'), and so the mechanisms which provide synonym-matching can be demonstrated. Finally, the structures which are created in the Population Database as a result of the linkages are moderately complicated. The major family grouping links together 17 persons and 4 conjugal families, while for one individual (Thomas ELDERS, P00147) there have been linkages across 11 nominal records.

16. As in earlier examples baptism and burial records are displayed in a generalised form as 'birth' and 'death' records.

17. Two additional items of information have also been included: these are the years of birth of Mary ELDERS (P00445) and Jane PORRIT (P00230). These items of information appear in the listing in Figure C.4, but were, in fact, obtained from records which contain no reference to the name 'ELDERS' or 'ELDER', and which for this reason do not appear in Figure C.1.

18. See Note 17.

19. This figure excludes the two references for which Thomas ELDERS is described as a father in marriage records (M00066 and M00068), since such references are currently ignored.

20. Close examination of the index does, however, reveal a slightly unusual item of information, and that is the year of birth of Mary ELDERS (P00445), viz. 1771. Since according to her marriage record (M00020) she was married as a spinster in 1826 this would imply that her age at marriage was 55. More detailed investigation does, indeed, confirm that this date of birth is in error and that the cause was an invalid link to an 1851 census record. This is further discussed in Section 12.2.3.

21. Between these two extremes one can also envisage various hybrid schemes, which would simultaneously involve both computer and manual analysis. Thus, for example, the computer could be employed to interrogate the total contents of the Population Database and to provide a report on any anomalies which it located: these could then be examined further by human intervention. Any family which contained more than, say, 8 children could, for example, be examined: a family of such a size could occur because two separate families had been incorrectly merged. Or, again, any family in which the wife was more than, say, five years older than her husband could be earmarked for further examination.

22. The subsequent analysis was carried out, in practice, by manually scanning through listings produced by the Population Analysis Subsystem and by making appropriate calculations. In principle, however, there is no reason why such an operation should not be
23. For this purpose it is considered appropriate to treat each group of synonyms as equivalent to the major surname.

24. The method adopted for calculating Student's $t$ is the 'small sample method' described in Moroney 1953, 231-3. Subsequent calculations also adopt this method, apart from the calculation at the beginning of Section 12.3.5, concerning the age at first marriage for grooms and brides. In this case the alternative method for paired samples is used (ibid, 227-31). In all calculations the stated confidence levels are as displayed in Table A 4 in Snedecor and Cochran 1967, 549.

25. However, the design of some linkage strategies can introduce a bias which favours the linkage of persons with uncommon names (Hershberg et al 1976A, 150-7). This arises because the strategies assign a higher confidence level to such linkages.

26. See Note 20.

27. As was mentioned in Section 11.2.1 the Population Analysis Subsystem currently provides no access to families via the Cohort Index, and consequently there is no simple way to obtain a display of all the families in the database. The method which was adopted for the present analysis was to obtain access via the Name Index, and this required the submission of an explicit 'SUBSET' command for each major surname in the database. Figure C.2 provides an example of the use of this method, in this case to display all the families for the surname 'ELDERS'.

28. This measure of the 'in view' duration represents the most restricted interpretation of the concept of being 'in view' that can be envisaged. A wider interpretation could perhaps be made to incorporate a retrospectively recorded event which was implicitly embedded in some other event. Thus, for example, each single post-1841 census event, because of the information about date and place of birth contained in it, effectively contains an embedded birth event: as such, it could be considered to possess its own 'in view' duration, spanning from the date of birth to the date of the census. However, since from the methodological viewpoint my central concern is with the difficulty of linking records for a mobile population it was felt that this difficulty would be brought into sharper focus by the use of the more restricted interpretation of 'in view'.

29. It should be added, however, that these figures are being presented in their most pessimistic form. In view of the relatively short study period there are some sizeable groups within the population for whom it is logically impossible to be in view for more than 24 years. Such a group consists of the 111 persons who were buried in the period 1813-37: this represents 7.1% of the whole population. And another, similar group consists of the 229 persons for whom there were baptisms in the period
1846-71: this represents a further 14.7% of the whole population.

30. 41 members of this group (i.e. 93%) have an age at death of 24 or less, and so, like the two groups described in Note 29 (and not necessarily disjoint from them), they cannot logically be in view for more than 24 years.

31. There is, of course, a secondary problem, and that is the slight change which is apparent in the methods of reporting over time.

32. But this is not guaranteed to remove all difficulties. See Note 31.

33. For the four father-son pairs in which both persons have a trade the trades of father and son are identical in two cases, viz. for a blacksmith and a carpenter. In the two other cases both fathers are blacksmiths and the occupations of the sons are shoemaker and engine fitter.

34. For the other two categories there is little data available, there being only two persons in the professional category and three in the trade category.

35. These results are somewhat higher than those reported by Wrigley and Schofield (1981, 255). In their case they found that for their sample group of 12 English parishes during the period 1800-49 the mean ages at first marriage for grooms and brides were 25.3 and 23.4, respectively, with a mean age difference of only 1.9 years.

36. The negative value at the low end of the range arises because estimated rather than actual dates of birth are employed in calculating the intervals.

37. Other properties of the two groups were analysed in an attempt to detect hidden biases, but no such biases appear to be present. Thus, the mean numbers of children in the farmer and labourer families are 6.2 and 6.1, respectively. And the corresponding mean time-intervals in years between the births of the first and last children for the two groups are 12.6 and 13.3, respectively.

38. Additional space was needed for the special directory used privately by IDMS for its overall supervisory and control functions. This required a fixed allocation of 1 Mbyte.
In this final chapter it is necessary to provide an assessment of what has been achieved from the present research investigation and to make some proposals about where future efforts might most fruitfully be directed.

At the beginning of the thesis I was at pains to raise a number of fundamental questions about the nature of computer problem-solving. I identified three major elements of this activity which I considered to be of particular relevance in the context of the present research. These are as follows:

1. The problem-solving activity which is enshrined in the present research demands a significant understanding of two quite separate disciplines, viz. historical demography and computer systems design. In terms of the historical demographic component it is necessary to understand the overall contribution and significance of record linkage to the research interests of demographers, as well as the more detailed issues concerning the nature of historical source materials, the methodological characteristics of the linkage process and the required properties of the end results of the linkage. In terms of the computing element of the research it will be clear from earlier chapters that significant computing
expertise is needed. In the first place it is necessary to have substantial competence in systems analysis and design, and the ability to organise a large software development involving complex, 'messy' problems. At a lower and more pragmatic level it is necessary to gain substantial competence in the use of a range of computing facilities, including in the present instance, the COBOL language and the IDMS database management system.

2. The solution of complex problems does not normally proceed in a rigidly sequential manner, with problem analysis being followed by solution implementation. Rather, the two activities are more likely to proceed in an iterative manner, with partial solutions leading to an increased understanding of the nature of the problem, which itself leads to improved solutions, and so on.

3. The overall strategy for problem-solving in the context of the present research is based on this concept of the inter-relationship between problem and solution. Central to the research is the objective of designing and implementing an operational record linkage system; this development is to provide a vehicle for achieving an intimate understanding of the nature of the problem. In retrospect, it is now quite clear, for example, that the significance and contribution of database management facilities to record linkage could not have been fully appreciated other than by exploring their use
in a practical situation.

At the beginning of the thesis I also identified my three main research objectives. Briefly, these were as follows:

1. to explore the use of novel computing strategies to achieve the family-based linkage of historical demographic data.

2. to produce a record linkage system embodying these strategies and capable of being applied to similar linkage problems by other researchers.

3. to present the material in this thesis in a form which makes it accessible to both the socio-demographic historian and the computer systems designer.

I am now in a position to take stock of the progress which I have made towards meeting my objectives, and I shall do this from within the framework of my perspectives about the overall nature of the problem-solving activity. I shall begin, therefore, in the first major section of this chapter by examining the problem of record linkage and the gain in understanding which has been achieved. In the second major section I shall turn my attention to the nature of the solutions to the problem: I shall consider both the solutions presented in this thesis and also, more tentatively, the kinds of approaches which may prove to be fruitful in the future. I shall then examine some of the wider implications of the research, and shall
discuss the relevance which nominal record linkage has for the controversial issue of data privacy. Finally, I shall summarise my main conclusions.

13.1 THE NATURE OF THE RECORD LINKAGE PROBLEM

One of the most significant fruits of the present research has been the increased insight which has been gained into the nature of the record linkage problem itself. This gain in insight has centred on the transition from the essentially person-based linkage viewpoint implicit in earlier linkage research to a more developed family-based concept, applicable to the linkage of both parish register and household census data.

Let us for a moment consider a comparable problem in computer software development, but one which is simpler and more fully understood, viz. the implementation of a compiler for a computer language, such as COBOL. Through a sustained analysis of the nature of this problem there has emerged a number of fundamental concepts: for example, the subdivision of the properties of a computer language into its lexical, syntactic and semantic components (1). And without the knowledge of such fundamental concepts the task of designing and developing a language compiler could prove to be extremely difficult.
It is therefore regarded as crucial that, in a similar fashion, an attempt should be made to identify the fundamental concepts of nominal record linkage if there is to be progress towards the development of viable linkage strategies. Unfortunately, because the linkage of historical records necessarily involves a host of 'real world' problems (2) the identification of fundamental concepts is made considerably more difficult than is the case for the problem of language compilation. Nevertheless, I shall proceed to isolate the more significant perspectives which have emerged in the present work, and move as far as I am able towards the identification of the desired fundamental concepts.

13.1.1 Problem Complexity

Nominal record linkage is a highly complex problem. This has been affirmed by others working in the field. Thus, according to Macfarlane: 'Despite excellent computing facilities and an enormous amount of effort and thought, we were finally unable to achieve automatic record linkage' (Macfarlane et al 1983, 20). And Hershberg, working with the person-based linkage of American data, has observed that 'While the model of record linkage is conceptually straightforward, its application to nineteenth-century census manuscripts for the construction of actual linked files is complex' (Hershberg et al 1976A, 140).
While certain parts of the problem, such as the need to handle name variations, are self-contained, reasonably well understood and capable of solution, others are more pervasive, relatively intangible and not readily amenable to solution. It is, of course, to be expected that in the early stages of exploring record linkage strategies the major focus will be the handling of the simpler, more tangible parts of the problem. This is understandable, since one is able to direct one's attention only at the problem as perceived. At a later stage, through prolonged contact with the problem and its solution, more intricate aspects of the problem are likely to surface, and one must then attempt to adapt one's viewpoint and development strategy accordingly. It must, however, be added that the inevitable 'shifting of ground' which is involved in such a situation presents a hostile climate in which to develop a computer system, especially where the underlying problem is complex. In the remainder of this section I shall proceed, therefore, to isolate for analysis some of the more difficult and unanticipated parts of the total problem.

From a systems analytic view the most perplexing characteristic of record linkage is the problem of defining it in a coherent and precise way. If, again for comparison, one considers the problem of language compilation one finds that it is not particularly difficult to provide a precise definition of this problem. Thus, once one is able to define the syntax and semantics of the computer language which is input to the compiler and the corresponding syntax and semantics of the machine instruction set which it produces then the problem of language compilation can be defined precisely in terms of the
implicit family structures of varying degrees of complexity. If a
systems analyst were invited to design a suitable input record format
for census data he would insist that such family structures should be
submitted in a much more regular and explicit form. In particular, he
would not be satisfied to leave all familial relationships expressed
solely in terms of each occupant's relationship to the nominated head
of the household. Instead, he would insist, for example, that each
spousal and parental relationship should be declared explicitly in
terms of the two people sharing the relationship. Without such a
measure of control over the input format the system designer is forced
to devise, as best he can, appropriate mechanisms for transforming the
arbitrarily structured records provided into a form which is more
compatible with the requirements of the main record linkage processes.

The other significant property of the file format for the
original source records concerns the information content. For a
language compiler there are strict syntactic and semantic rules
concerning how statements in the language are to be constructed and
ordered. And where a programmer breaks any of the rules the compiler
is required only to provide appropriate error diagnostics, i.e. there
is a requirement for it to complete the compilation only where all
statements have been entered correctly. By contrast, the position
with regard to the information content of the original source records
in record linkage is that there are no strict rules concerning how the
information is to be constructed and ordered. Any field may be left
empty, and for any type of information, e.g. surnames, there are
guaranteed to be inconsistent and ambiguous forms of presentation.
conversions which need to take place between the language and the instruction set (3). In other words, the problem of compilation can be completely defined in terms only of two file formats and the transformations which are required to produce one from the other.

Superficially, it would seem that the same method of problem analysis should be applicable to record linkage. Thus, one would begin by attempting to define two file formats, one describing the structure of the original source records and the other the structure of the linked data. One would then seek to define the problem of record linkage in terms of the transformations needed to produce the second file from the first. While this intuitively appears to be a fruitful way in which to conduct the problem analysis, in practice it is fraught with difficulties.

Let us first consider the task of defining the file format for the original source records. The most notable property of this format is that it has an arbitrary structure and content, i.e. one which is completely dependent on the types of source records which are to be linked and their intrinsic information content and structure. For most computing problems, including that of language compilation, the input file format and content is non-arbitrary: indeed, it is usually designed so as to facilitate the required computational processes. While some source records, such as post-1812 burials, may have a moderately regular structure others can be exceedingly complex. Household census data, for example, although possessing a disarmingly simple explicit household structure, can contain a vast range of
Furthermore, far from the system being able to flag such occurrences as errors it is required to adopt strategies which will make appropriate corrections and compensations for the information deficiencies at all stages of the processing. This requirement alone must cause the system to be at least an order of magnitude more complicated than it would otherwise be.

Let us next consider the task of defining the file format for the linked data. At its simplest this format must be capable of subsuming all the individual elements of the source record types which are present in the input file format, but also it must be able to encompass the various 'regroupings' which will have taken place during record linkage. For person-based linkage there may be fairly limited regroupings, involving, for example, the simultaneous ordering of people into nominal, cohort and occupational groups. But for family-based linkage, regroupings can occur at additional levels, viz. children to their families of origin, parents to their families of procreation and husbands to wives. As a consequence the resulting file format is inevitably very complicated (4).

In addition, some of the deficiencies which are present in the file format for the original source records must of necessity be transmitted to the file format for the linked data. For example, wherever it is necessary to allow for a missing data item in an original source record there must be a corresponding allowance for a missing item in the linked data. And this problem of missing data is further exacerbated by the fact that for each individual there will be
complete source records missing, because of deficiencies in the recording process and because of personal mobility. The net effect of all these factors is that in a file of linked data the information held about any one person can be in a potentially very large number of distinct states. For example, there may or may not be information about a person's date and place of birth, date and place of death, identification of parents, spouse(s) and children, and so on. Additionally, for a woman the maiden name may or may not be present. This variability in the contents makes the actual file structures (and therefore also the associated processing) potentially very complicated. If one were designing a 'modern' population database, by contrast, one would insist on drastically limiting the number of distinct states in which the information about any one person would need to be maintained. One would insist, for example, on always holding a person's date and place of birth and for a woman the maiden name. The result would be that the task of handling the information in such a database would be considerably simplified.

It should be clear that the task of defining input and output file formats for record linkage is complex. From this it follows that there can be no convenient way of defining the problem of record linkage in terms of the transformations which are needed to produce the output file from the input file. My conclusion must therefore be that given this perception of the problem there can be no completely systematic way of developing a solution.
I shall now briefly address some of the more detailed problem areas where the scale of the problems was to some extent unanticipated. Firstly, let us consider the task of linking post-1841 census data, since this has presented one of the most significant challenges of the entire project. As I noted earlier, each census household effectively contains an arbitrary number of arbitrary pieces of genealogical structure, and in the process of record linkage these must be linked to other such pieces, as well as to other types of records. This is a particularly complex problem. Although the 'relationship to head' information does provide an explicit link between the individual members of the household and the head, there are severe difficulties in making use of it.

In the first place, one is required to extrapolate from the 'relationship to head' information to establish what relationships exist between each occupant and every other person in the household. In some circumstances this is a relatively simple operation: for example, where the presence of two 'son' relationships would indicate a sibling relationship. But there is a whole range of much more complex relationship-pair situations. For example, if a 'sister' and a 'nephew' are encountered in the household then one must try to establish whether a parent-child relationship exists between them. This decision must be based on an examination of the other available information about the two people, such as surname, marital status and age.
Secondly, certain designations, such as 'grandfather', 'uncle' and 'cousin', are particularly difficult to use because of their imprecision. Thus, in each case it is uncertain whether the relationship exists via the head's father or mother. And even in an apparently simple household consisting of a 'head', 'wife' and 'son' there can be no guarantee that the son is not a stepson of the wife. In some circumstances the presence of additional family members can serve to reduce the uncertainty, but the provision of the necessary algorithms to handle the large range of familial situations which arise would be a formidable task.

Finally, for those people who are not related to the head one should ideally attempt to deduce and make use of any family relationships which may implicitly exist between them. If, for example, one were to locate a married male and a married female living in the same household, both sharing the same surname and having similar ages, one might conclude that they were husband and wife. Sibling and parent relationships could be deduced in a similar fashion (5).

In view of these difficulties it should be clear that the task of exploiting fully all the information present in each census household and subjecting it to the processes of record linkage demands a strategy of considerable sophistication and complexity.

In this review of detailed problem areas I shall finally refer again to the difficulties of handling spatial and occupational
information. As we saw in Chapter 3 there is a whole range of problems. Placenames and occupational titles share, with personal names, the problems of synonymy and homonymy. But, in addition, they suffer from variable precision and boundary and identity problems. In the present implementation the associated facilities are relatively unsophisticated, but it is clear that the task of providing a more comprehensive range of facilities would present significant conceptual and organisational problems.

Firstly, let us consider the difficulties of handling spatial information in a comprehensive manner. In addition to the facilities already provided one would ideally wish to be able to satisfy two further requirements. In the first place one would wish to be able to associate an Ordnance Survey grid reference with each placename. As a result one would then be able to introduce the concept of 'distance' into the Population Database, and so provide a convenient, automated means for studying migration (6). And in the second place one would also wish to have some means for organising places into a geographical hierarchy: for example, relating placenames to particular counties, counties to countries, and so on. Provided with such facilities one would then be able to locate, say, all the Irish-born people in the Population Database. Unfortunately, primarily because of the problem of homonymy, i.e. the occurrence of several places with the same name, the task of providing such facilities would be difficult.

Similar difficulties would be encountered in the attempt to handle occupational information in a more comprehensive manner. In
addition to the facilities currently provided one would wish to be able to satisfy a number of other requirements. In the first place it would be desirable to be able to make more effective use of the detailed occupational descriptions which are provided in the census. At present there are no mechanisms for handling the attributes relating to the area of land owned, the number of people employed and, for those in trades, the employment status, i.e. whether apprentice, journeyman, master or retired. In the second place it would be desirable to have some means for organising occupations into one or more hierarchical classification schemes: for example, each occupation could be related to its position in an industrial classification and also to its position in a social class hierarchy (7). Provided with such facilities one would then be able to locate, say, all those present in the Population Database who had at some time been employed in agriculture or, again, those who were in professional and managerial employment. However, it should be added that, as for the organisation of spatial information, the problem of homonymy makes the task of providing appropriate facilities difficult.

From the present discussion it should be clear that nominal record linkage is a complex, multi-faceted problem. And as already indicated, because of the complexity of the input and output file formats there appears to be no completely systematic way of developing a solution. In addition, one is confronted with a whole range of problems associated with the linkage process itself, and one is also confronted with significant problems which relate to the form and content of the individual types of records. In particular, census
records present severe organisational difficulties, especially where one's objective is to confront the more complicated technical requirements of family-based record linkage. Finally, there is a quite separate range of problems concerning the methods of organising certain types of information and the way in which 'meaning' is to be attached to them.

13.1.2 An 'Architectural' Perspective

One of the more significant changes in perspective which have occurred during the present work involves the transition from a concern with 'algorithm' to a concern with 'architecture'. Expressing this in a slightly different, and more computer-oriented, way one could say that the concern has moved from 'program' to 'data'. In retrospect, one can see that this change in perspective came about largely in response to the discipline of confronting the problems of family-based record linkage.

When confronted with the simpler problem of person-based record linkage the characteristic response has been to view it primarily as a problem of 'mechanism', i.e. how to construct appropriate programs to accomplish the required inter-record linkages. From this perspective one's concern is to locate or else devise the necessary algorithms which will contribute to the overall solution. Thus, for example, Soundex and record-matching algorithms are typically viewed as the key components in any solution and represent the kinds of topics which are
considered to be of major concern. With the emphasis being placed on 'mechanism' the consequence is that the characteristics of the input and output data are accorded a secondary significance. In short, it is the program which is seen to be at the centre and which has importance; the data which passes through it is viewed as a transient and of less significance.

This orientation is understandable when one considers the characteristics of person-based record linkage. In general, both the input and output data formats possess a relatively simple structure. The input records of a particular type usually have a regular structure, and the output data frequently consists of a sequential merging of input records into simple personal dossiers. By contrast, the programs which carry out the linkage can have considerable inherent complexity and so they naturally become the main focus of interest and attention.

The change in perspective which has occurred within the current project has precisely reversed the roles accorded to program and data. In the modified view it is the data which is seen to be at the centre and which has importance; the program is viewed as the transient, whose sole function is to 'enter', transform the data in some way and then 'exit'. This shift in emphasis is clearly the result of the substantial increase in the level of complexity of the data when we progress from person-based to family-based linkage. One must now deal, not only with the implicit complexities of familial relationships in household census data, but also with an end product
of the linkage which possesses a considerably more complex structure. Since database management systems are expressly designed to assist the handling of complicated organisations of data, it is commonly the case that users of database systems will naturally come to place increased emphasis on data at the expense of programs.

With the adoption of this new perspective the emphasis now rests on the design of data structures, and the examination of input and linked data file formats in the previous section exemplifies this orientation.

This perspective has had a subsequent crucial impact on the progress of the later phases of the 1851 Census National Sample Project (Anderson and Welford 1984; 1988). The entire conceptual and strategic approach to the organisation of family information in this project and, more generally, the development of a 'public data' distribution format grew directly from the architectural perspective enshrined in the present research (8). Preliminary analyses of data generated in this public data format has recently been published (Anderson 1988).

By applying an architectural perspective to the problem of record linkage in a rather more rigorous fashion one might then aim to analyse this problem, not primarily by considering mechanisms, but rather by attempting to design a range of data structures, including those for the input and linked data files, but also including a number of 'intermediate states between these. Only after completing such a
comprehensive 'data analysis' would one then move on to consider the individual 'mechanisms' (algorithms, programs, etc.) which would be needed to provide the required transformations between the various 'data structure states' (9).

This new perspective is described as 'architectural' in the sense that it is concerned primarily with the overall design of the data structures which are required, rather than with how they will need to be manipulated. This orientation firmly exemplifies the 'top-down' approach, which was introduced in Section 4.1.3, although in the present context the design centres on data, rather than program, organisation. It is clear that further progress in the development of family-based record linkage techniques will demand an increased emphasis on data organisation issues and a corresponding reduction in the relative attention given to program algorithms.

13.2 SOLUTIONS TO THE RECORD LINKAGE PROBLEM

13.2.1 The Current Solution

I shall now provide an evaluation of the linkage strategies which have been developed as an integral part of the present research. And as a preliminary to this I shall first comment on the overall development strategy which has been adopted.
Throughout this thesis I have drawn attention to my use of evolutionary and 'top-down' development strategies. In view of the perceived complexities of family-based nominal record linkage it is considered that the use of such strategies has been of crucial importance in enabling a working system to be produced. At the same time the ability to employ this system as a suitable 'test-bed' for the exploration of linkage problems has been considerably enhanced by its strict modular design. And were it desired to develop the system further, for example, to handle 1841 census data or to provide a relational database interface, then it is considered that such enhancements could be integrated into the present design with minimal difficulty.

Implicit in the evolutionary approach is a recognition that systems development is best organised by subdividing the total task into a number of discrete stages, and that one should aim to develop primitive mechanisms initially and then extend them progressively with additional facilities. The particular benefit of this approach in a research context is the strategic flexibility which it permits, enabling design options to be kept open for as long as possible. The danger from adopting alternative, 'monolithic', design approaches is that it becomes much harder to reverse early design decisions, and so the scope for exploration and experimentation is considerably reduced.

These considerations should be borne in mind in relation to my subsequent assessment of the present linkage strategies. Thus, the strategies which are embodied in the current system should not, in any
sense, be regarded as some 'final design', but only the position reached after an arbitrary number of design iterations. The need for enhancements and refinements is therefore freely acknowledged.

In the remainder of this section I shall evaluate the linkage strategies which have been developed. I shall do this by first considering those characteristics which are thought to have been particularly effective, and I shall then proceed to analyse some limitations in the present approach.

Let us initially view the record linkage system as a 'black box', and consider its external properties, rather than its internal structure and mechanisms. A fundamental requirement of any record linkage system is that it should be straightforward to use and that the facilities for entering the records should ensure that minimal errors are introduced. In the present system, therefore, the interface which is provided for the user is convenient and flexible. The primary source records are presented to the system in a form which closely matches their original format, and so the user is not required to reformat and transpose fields, operations which could introduce errors. The records after input are subjected to a comprehensive range of data validation procedures, and this further guarantees that a high level of data integrity will be achieved.

The method of priming the name directories is also convenient, and the user is thereby given considerable and simple control over the handling of all names. Furthermore, this strategy for resolving the
problems of name variations is an effective and reliable one, and it possesses significant advantages over alternative, algorithmic Soundex-type schemes. These advantages were extensively examined in Chapter 3. It should be noted, in particular, that the use of a directory permits names which do not sound the same, such as 'John' and 'Jack', to be treated as synonyms: Soundex is not designed to cope with this. The extra control over names which a directory provides can also facilitate the detection of data preparation errors: Soundex algorithms normally lack this capability, as they are designed to accept as valid any alphabetic string which is submitted. An additional significant advantage of the directory approach is exemplified in the case of the Christian name directory by the facility which allows user-specified gender information to be associated with each name (see Section 3.1.2). Such a facility, which is incompatible with the Soundex approach, is valuable in that it permits the carrying out of consistency checks on the nominal data which is presented for linkage. A more general comment is that this directory approach, based as it is on the submission of user-supplied 'expertise', is consistent with the recent developments in the field of intelligent knowledge-based systems.

At the record linkage stage it is desirable that the user should have a measure of control over the operations which are carried out. In the present system, therefore, he is provided with commands which enable him to select for linkage records of restricted types and/or for particular names or groups of names. The ability to have displayed those records which have participated in a particular
linkage operation also provides the user with a convenient means of monitoring the validity of the linkage strategies.

The strategies as currently developed achieve the family-based record linkage of nineteenth century census and parish register data. The analyses carried out in Chapter 12 demonstrate that the particular strategies used are reliable, with few linkage errors being encountered.

The whole purpose of carrying out record linkage is to provide the user with an enriched data source, and so the provision of appropriate methods of accessing the linked data is of paramount importance. In the present system this is achieved by providing the user with a wide range of controls enabling him to obtain selective displays of the linked data. Using these he may request details about various groups of persons and families in the population. Thus, persons may be indexed via their birthdate/birthplace, nominal and occupational classifications; families may be indexed via their nominal classifications. For each person or family selected the user may choose to have information displayed at a range of levels of detail. Observation of such displays enables the user to monitor the 'correctness' of the linkage and carry out analyses. An important feature of these displays is the comprehensive use of unique record identifiers for all source records, persons and families. Such identifiers facilitate the cross-referencing of pre-linkage and post-linkage records and also the tracing of familial relationships in the Population Database.
The total record linkage operation subdivides conveniently into several discrete and simple stages, enabling the user to approach the task in a controlled manner and to ensure that each stage is accomplished correctly before proceeding further. In its operation the system is efficient, reliable and robust. Finally, the fact that the system has been implemented using COBOL and IDMS, both of which are widely available, should ensure that it can be made operational on many available mainframe machines.

Let us now turn our attention to the interior of the 'black box', i.e. the internal structure and mechanisms of the system. The use of a database management system has imposed a distinct discipline on the way in which the data is structured and manipulated. This has been especially beneficial in promoting the 'architectural' perspective described in Section 13.1.2. In terms of the usage of IDMS one is constrained initially to consider the organisation of one's data rather than one's programs. Thus, one begins by designing the schema, a complete description of one's database in terms of the areas, records and sets of which it is composed. Only then does one begin to consider the programming implications. So in the current implementation the data structures present in the Directory, Source and Population Databases constitute the fundamental design features around which the programming design is organised. Each of the subsystem programs effectively operates as a 'transformer' between one data structure organisation and another. And the division of the total program into separate subsystems and modules disaggregates the highly complex processes involved in record linkage into a more
manageable series of sub-processes and operations.

One of the most significant data transformations involves the conversion of primary source records into internal coded equivalents. This 'preprocessing' function converts the data into a form which facilitates the subsequent manipulation. The problems which are to a great extent filtered out at this stage are as follows:

- the handling of free format text and the use of dittos.

- variations in the forms of entry for particular record types, e.g. for pre- and post-1837 marriage records. The setting up, for example, of a single, all-purpose internal marriage record format in place of the corresponding range of data collection formats is considered to be a particularly important design feature.

- name variations.

- variations in presentation of information relating to ages and dates.

The other major transformation occurs when data is transferred from the Source Database to the Population Database. The internal structure of the Source Database reflects the input requirements of the Record Linkage Subsystem. Thus, it essentially consists of isolated records, together with a surname index. By contrast, the
Population Database, which is much more complex, is intended to model the 'real world' structures of a population. It therefore contains entities such as persons, families, households, events, and so on, together with associated entities for providing multiple access routes to persons and families which may be of interest.

This clear subdivision into a separate Source Database and Population Database, each with its own characteristics, represents one of the most distinctive features of the current design. The various operations which precede the loading of records into the Source Database effectively filter out many of the tactical and organisational problems which must be handled. It is subsequently left to the Record Linkage Subsystem, in moving data to the Population Database, to deal with the remaining problems. This disaggregation of function is a crucial feature, serving to make manageable what would otherwise have been extremely complicated.

The final and most important feature of the current design is the method of modelling genealogical structures and life-histories in the Population Database. The use of a 'PERSON' record, a 'FAMILY' record and three interconnecting sets, as established in Section 4.3.3, provides an extremely powerful modelling device (10). Genealogical structures consist of rich networks of familial interconnections, and there is no convenient method of handling them using conventional programming and hierarchical filing techniques. However, the facilities provided by IDMS and the use of the simple 'PERSON'-'FAMILY' record and set organisation have provided an elegant
and completely general solution. In spite of its apparent simplicity this arrangement has been able to support satisfactorily and conveniently the whole range of genealogical structures which it has been necessary to consider. Thus, in addition to the elementary family relationships it can readily handle sibling relationships in the absence of parents, in-law relationships, single parent families and remarriage and step relationships. Furthermore, the dynamic building up of life-histories by means of 'event' records and an associated set, 'P-LIFE-EVENT-SET', as illustrated in Figure 9.5, has proved to be a most effective design feature.

While the present system largely satisfies the functional requirements which were originally proposed for it, there are some features of its design where further enhancement can be envisaged. I shall now briefly examine these.

Although the present linkage strategies have been designed to process particular kinds of nineteenth century records, the underlying system mechanisms enshrine a high degree of generality. The enhancement of the system to handle additional record types would therefore be a moderately straightforward task. At the same time one would wish to increase the level of generality even further by incorporating the concepts of person and family occurrence templates, as presented in Section 10.3.2.

While the present system design permits the user to intervene considerably in the total record linkage process, one can envisage
that additional levels of interaction could be incorporated. For example, it would be possible to allow the user to specify dynamically certain parameters of the linkage process (e.g. the age differences which are permitted in particular record matching situations). It would then be possible in theory for the user, by adjusting these parameters, to 'fine tune' the system to achieve optimum linkage results. Additionally, the user could be provided with facilities for overriding linkage decisions made by the system. One can envisage, for example, a situation where the user has observed that the system has created two persons called 'John SMITH', where in fact it is judged that there should only be one. Facilities which would enable the user to instruct the system to merge the two persons would clearly be useful. In addition to such facilities one can also envisage a requirement for facilities which would have the opposite effect, i.e. to cause a splitting of a person for whom records originating from different people had been incorrectly merged. The application of these ideas has recently been exploited within the context of the 1851 Census National Sample Project (11).

Finally, it would be worthwhile to introduce some enhancement to the facilities for analysing the contents of the Population Database. In view of the absence of the required mechanisms within the IDMS database system the most fruitful approach would be to provide a mechanism for creating an 'export' file in a format suitable for reading by an appropriate analysis package or database system (12).
13.2.2 Future Approaches

It is necessary now to make some proposals about how future efforts might most fruitfully be directed in order to develop improved methods for linking records. I shall initially restrict my attention to a consideration of what might be achieved in the short term, and here my thinking will naturally relate to the kinds of enhancements to the current approach which were identified above.

A characteristic of the present approach is, perhaps, that it addresses too large a problem. And so one potentially fruitful strategy would be to actively seek ways of reducing the scale of the problem. The use of a 'post-processor', such as an analysis package or relational database system, for analysing the contents of the Population Database exemplifies this approach. And quite clearly the effort needed to produce the required 'export' file would be significantly less than that needed to implement a range of special-purpose analytical facilities.

One might seek to achieve a similar reduction in the scale of the problem at the record input stage. Here one might envisage the use of a separate preprocessor which would be responsible for converting the data from its initial, free format state into a fixed format state, and one which was much more easily handled by the system (13). Considering this problem in a wider context, it is obvious that it is a costly overhead for all users of primary historical source materials to have to carry out a conversion such as this as a preliminary to
their data analysis and manipulation activities. How much more convenient it would be if the data were to be preprocessed once and for all at source and then distributed in a simple, fixed format state (14).

The other significant area where some progress might be made in the short term is in the further generalisation of the record linkage strategies. Proposals for the use of person occurrence and family occurrence templates were discussed in Section 10.3.2. The main purpose of these proposals would be to establish person and family matching mechanisms which were entirely source type independent. This would make the overall linkage strategy much simpler to implement and refine. It is believed that this kind of approach would also prepare the way for a more rule-based orientation.

In the longer term it will be necessary to assess and, where appropriate, make use of the new computing facilities which become available. The development of database facilities in the past decade has been of enormous relevance as a contributor to the solution of the record linkage problem. In particular, it has made manageable the handling of the complex network structures which are involved in family-based linkage. During the next decade, however, we may expect to see the development of other techniques which will assist the further generalisation of the record linkage process.

A major, relevant subject of interest at present is the potential for exploiting the concepts of artificial intelligence in the next
generation of computers, the so-called 'fifth generation'. 'Fifth
generation computers will think like humans. They will draw
inferences from a series of known facts about the world.' (Bird 1982,
19) If, indeed, the next generation of computers does succeed in
fulfilling such prophesies then it is quite certain that this will
require the problems of record linkage to be examined from a modified
perspective. And since in artificial intelligence there is an
emphasis on the use of rule-based techniques it is probable that
future developments in this area will profoundly assist the production
of generalised linkage strategies. The recent work of the RESEDA
project in France, for example, involving the application of the
techniques of artificial intelligence to historical research, does
give an indication of the shape that such developments might
take (15).

Some recent work with modern demographic records has demonstrated
what can be achieved from successful record linkage, and this will, no
doubt, act as a spur to further progress with the linkage of
historical records. This is the work of the OPCS Longitudinal Study
(OPCS 1982; 1985), and it has involved the manual linkage of 1971
census data to a range of vital records, such as births and deaths,
for a 1% sample of the population of England and Wales. Although the
study has restricted itself essentially to person-based linkage, it
has, for example, been able to relate mortality statistics to a whole
range of social and demographic determinants, including educational
qualifications, economic activity, household and housing
characteristics, area of residence, etc. From its analyses it has,
for example, demonstrated that mortality is higher for those who live in council houses than for owner-occupiers (OPCS 1982, 66) and that it is higher for the unemployed than for those in employment (op cit, 195). Such a study demonstrates very powerfully the scope of what can be achieved when rich data sources are brought together, albeit by manual means, and it has recently been described as the most important development which has taken place in record linkage in Britain since 1962 (Baldwin et al 1987, 8).

Meanwhile in the United States the Amoskeag study has in a similar fashion carried out the manual linkage of a whole range of employee, vital and taped interview records (Hareven 1982). This large-scale study has enabled the family and work patterns of people living in an American industrial city in the early part of this century to be explored in depth, and it once again demonstrates the power of record linkage.

Returning now to the subject of automated record linkage it may be prudent to conclude on a cautious note. The provision of new and more appropriate computing facilities is, of course, to be welcomed. But such facilities cannot of themselves be guaranteed to solve the intricate and highly complex problems of historical, family-based record linkage. Let us, again for comparison, consider the use of the IDMS database system in the current work. This system has proved to be of major, and perhaps crucial, assistance in the achievement of a workable solution. But even with this assistance the resulting linkage system, viewed as a database application, must be regarded as
very complicated (16). It is therefore to be expected that even with
the provision of the kinds of rule-based techniques which are becoming
available, it will not be possible to implement a viable,
general-purpose, family-based record linkage system without a more
profound understanding of the methodological problems, the ability to
apply the necessary design skills and also considerable effort.

13.3 SOME WIDER IMPLICATIONS OF THE RESEARCH

In Chapter 1 I suggested that the findings of this thesis might
not be restricted solely to issues concerned with the specific
interdisciplinary subject under investigation, but that the overall
approach adopted and lessons learnt might have relevance to other
researchers working in quite disparate application areas. With the
rapid development of computing technology, particularly into areas
such as artificial intelligence and computer-aided learning, people in
non-computing disciplines are frequently struggling with difficulty to
understand the relevance of this technology in interdisciplinary areas
where there may be no existing body of experience or expertise to
guide them. It is with such people in mind that the following
observations, derived from the experience of the current research, are
made:

1. It is necessary to avoid being rushed into adopting premature
   computing solutions to one's substantive problems. The
appropriate technology for a particular problem may not, in fact, yet exist. It should be remembered also that some problems which appear simple for a human being (e.g. face recognition) are extremely complicated or even impossible for a computer.

2. Where one is adopting a computer solution, one should try to select the most appropriate technology for the particular problem. Firstly, there is the danger of acting too conservatively and choosing to use technology which has served well in the past, but which may be inappropriate for one's new, rather different kind of problem. The danger at the other extreme is 'to jump on the band-wagon' and choose some new kind of technology merely because of its novelty value.

3. It should be borne in mind that when engaged in research it can be advantageous to employ computing technology which is well established, rather than that which is experimental and 'state-of-the-art'. Well established systems will normally be comparatively error-free, as well as being fully documented and supported: experimental systems can present significant risks, especially where the support arrangements and future availability are in doubt.

4. It should be recognised that computer projects require people with significant and appropriate technical skills. The ability to write programs of one or two hundred lines will not
normally be a satisfactory qualification for someone who is required, for example, to design and implement a complex database application.

5. A sufficient amount of time and effort should be devoted to understanding thoroughly the nature of one's problem before a solution is sought. There is often an understandable pressure to make progress and generate tangible results, but if this pressure is too persistent there is a danger that solutions to partial or even the wrong problems will emerge.

6. At all times the aim should be to maintain a global view of the total problem and a clear picture of the end product to be achieved. This should deter one from getting too side-tracked into peripheral, but possibly interesting, areas of the problem.

7. Where significant amounts of data are to be prepared, particular attention should be devoted to establishing the appropriate format for this. The success of an entire enterprise can depend on the precise form, content and accuracy of the data. As far as possible the data preparation stage should be delayed as long as possible, and at least until it is clear that the proposed data input formats will be wholly appropriate for one's purposes. Once data preparation has started in earnest it will be found that many design options will quickly disappear.
8. There should be an encouragement to use the kinds of disaggregative implementation strategies described in this thesis. The entire project should be divided into a series of small steps, so that at any time one will be working towards limited, achievable goals.

9. Top-down, modular programming methods should be applied, so that the resulting programs will be comprehensible, reliable and amenable to modification.

10. Overall, one needs to be aware that time and resource estimates for computer projects are often wildly optimistic. Periodically one should be reminded of the old computer adage: 'If you think a problem is simple it will turn out to be difficult; if you think a problem is difficult it will turn out to be impossible!'

The application of the prescriptions described above had a direct and important relevance to the progress of a project with which I was involved during the period 1980-5. This was the LAMSAC-initiated SASPAC Project (17), whose purpose was to develop a highly portable computing system for handling the 1981 Population Census Small Area Statistics data for Great Britain (Welford and Stott 1983). I was the chief systems designer of SASPAC, and in this project I promoted most of the systems approaches developed in the current research, including the use of a strict top-down, modular design methodology.
An important by-product of the project was the design of a highly portable magnetic tape data interchange format for the census data, which was subsequently used by OPCS to distribute the data to all customers (Rhind 1983, 366). The design of this data format was influenced considerably by the ideas about public data formats developed in the current research. Such has been the success of this method of distribution that OPCS has adopted the same file format design principles for other data which it distributes (Wright 1983).

SASPAC was a joint winner of the British Computer Society's 1983 Social Benefit Award and is in use at more than 150 local and national government computer installations.

13.4 RECORD LINKAGE AND DATA PRIVACY

A thesis on nominal record linkage may be considered incomplete unless at least some comment is made about what can be regarded as the converse problem, viz. that of data privacy. I shall now therefore briefly address this matter.

The concern which is commonly expressed in relation to the issue of data privacy rests on the misuse to which personal data might be put. In particular, there is the fear that:
1. personal data which has been collected for one purpose may be used for some other purpose.

2. data from disparate sources may be linked to form personal dossiers.

On reflection it will be clear that the present research has required a fundamental involvement in both of these activities, albeit with data for a historical population. Inevitably this involvement has caused me to consider the possible implications for data privacy and to become concerned with the social repercussions which might arise from the linkage of modern records. At the same time some attempt has been made to identify technical ways in which the privacy of the individual might be protected (18).

Since record linkage is such a complex task for historical populations this might curiously suggest to us that the very characteristics which are responsible for the linkage problems could be turned to advantage in our present situation to enlarge personal privacy. Thus, what this has led me to propose is the deliberate adoption of 'regressive' design techniques. It is clear that the best way to facilitate the record linkage process is to give every person a unique person identifier and to include this in each of his nominal records (19). Therefore, it follows that in order to safeguard personal privacy we should adopt the opposite and more regressive approach of allowing each person to have several alias identifiers, thereby making record linkage much more difficult. Such alias
Identifiers might, for example, include a person's National Health Service Number, passport number, bank account number, etc.

A second way in which a regressive design approach could be used to increase the potential for personal privacy would be to set out deliberately to hold nominal records in a disaggregated or 'delinked' form, with the separate parts stored in physically separate locations. Thus, for example, medical records could be held in such a 'delinked' form, with the main medical data being held on a central computer, but with the key identifying information (e.g. name, address, etc.) being stored on the appropriate general practitioner's local microcomputer. By employing this arrangement the 'anonymous' medical data on the central computer could be exploited to the full for analytical and other purposes, while the privacy of each individual would still be reliably maintained (20).
13.5 CONCLUSIONS

My main conclusions may be summarised as follows:

1. The family-based nominal record linkage of historical data is a multi-faceted problem of significant complexity. From a systems analytic view the most perplexing characteristic of the problem is the difficulty of defining it in a coherent and precise way.

2. In addition to the well understood, 'visible' problems, such as that of handling name variations, there are many difficult 'hidden' problems. One of the most significant of these concerns the immense difficulties which can be involved in the development of computer programs to handle awkward 'real world' data. Another major problem concerns the handling of the enormous range of family structures which are implicitly contained in household census data.

3. For the family-based linkage of nineteenth century demographic data for small communities the problem of organising the relatively complex, information-rich source records is more significant than the problem of nominal ambiguity.

4. Given the significant complexity of the problem a sound strategy for achieving a solution needs to be based on an iterative overall development plan, with an understanding of
the problem being progressively gained through the attempt to
develop primitive, partial solutions.

5. The task of producing the overall design for a record linkage
system requires the adoption of an 'architectural', rather
than an 'algorithmic' perspective. From the 'architectural'
perspective the prime objective is considered to be the design
of a number of discrete 'data structure states'. The design
of the 'mechanisms' which are needed to provide the
transformations between these states is deemed to be of
secondary and subservient importance.

6. Database management facilities have a crucial role to play in
solving the problems of nominal record linkage. They are
important not only for their ability to control the
organisation of records and the connections between them, but
also because, at a higher level, they can provide a powerful
tool for modelling complex family and other structures.

7. The facilities of a CODASYL-type database system, such as
IDMS, are well matched to the problems of linking nominal
records and organising genealogical and life-history
structures. For the post-linkage, analytic requirements a
relational database system is more appropriate.

8. The use of name directories provides a much more effective and
reliable method of handling the problems of name variations
than alternative algorithmic, Soundex-type schemes.

9. The linkage of nineteenth century demographic data for small communities can be successfully achieved by employing a 'sequential' strategy, involving the prior linkage of the more information-rich records.

10. The problem of developing more generalised family-based record linkage strategies, e.g. to handle a wider range of input record types, is conceptually and organisationally complex. It will require the use of novel, rule-based computing 'mechanisms' and an associated refinement of our understanding of the nature of the problem.

11. The task of solving the problems of nominal record linkage will continue to demand significant interdisciplinary expertise. It will require not only an intimate and profound awareness of the methodological issues, but also the necessary systems analysis and design skills needed to develop complex software systems.
NOTES

1. For a lucid description of the fundamental concepts of language compilation see Welsh and McKeag 1980, 37-228.

2. These problems were extensively discussed in Chapter 3.

3. It is assumed in this discussion that a non-interpretive compiler is being considered, i.e. the type of compiler which produces machine object code, rather than an intermediate code requiring interpretation.

4. Thus, for example, in the IDMS implementation illustrated in Figure 9.5 it has been necessary to employ seventeen record types and seventeen set types to provide the required regroupings in the linked data. And this particular arrangement does not by any means represent a completely generalised and exhaustive population database model.

5. Such facilities as these were in fact recently developed for use with the census data constituting the 1851 Census National Sample, a project which I co-directed during its later phases with Professor Michael Anderson at the University of Edinburgh (Anderson and Welford 1984; 1988).

6. I recently developed such facilities for the 1851 Census National Sample Project (see Note 5).

7. A detailed examination of these issues is provided in Armstrong 1972.

8. The 'public data format' for the census data represents, what one might call, an optimal data structure distribution format. As such, it is designed to facilitate the transmission of the data to the user and also simplify its subsequent handling and analysis. Thus, the emphasis is on the establishment of a comprehensive data format design, rather than the provision of sophisticated analytical tools.

   A significant design feature of the public data format is the incorporation of family records. These records are only implicit in the original data as collected, but are expressly manufactured for inclusion in the distributed data. The concept and design of the family records, and the strategies for producing them, have all developed from the corresponding concepts and mechanisms employed in the present project.

   It is of note that a comparable, modern data source, viz. the General Household Survey, has been distributed by OPCS in a form which makes it highly inaccessible to users. Thus: 'It is now widely accepted that the General Household Survey provides a very valuable source of data for secondary analysis by social science researchers, but that the complexity of the data structure
presents severe problems for the average researcher' (Arber et al 1981, 1). One significant deficiency in the data structure is that, while it facilitates analyses based on the 'household' and 'person', the absence of an explicit 'family' entity makes it difficult for family-based analyses to be carried out. A project at the University of Surrey has, however, developed software to overcome such weaknesses in the distributed format (Gilbert et al 1980; 1982). Thus, for example, it has developed a flat file distribution format in which more than 60 special variables are added to each person record to provide 'cross-linking' to other individuals in the same household or family unit. OPCS has subsequently introduced a distribution format for the General Household Survey which enables it to be loaded directly into the SIR (Scientific Information Retrieval) data management package (Dale et al 1988, 89-90). As such, it can facilitate the carrying out of family-based analyses.

9. Gane and Sarson 1979 provides a comprehensive introduction to a range of data-oriented techniques which can assist this kind of approach to system development. In particular, the introduction of an associated graphic notation, viz. the data flow diagram, represents a welcome development. Essentially this notation provides the data equivalent of a program flowchart.

10. These modelling concepts were subsequently employed in the design of the public data format for the 1851 Census National Sample Project (see Note 8).

11. See Notes 5 and 8. During the production of family records in the public data format file the user is able to submit so-called 'forced relationship' commands. These enable him to manipulate the familial links between the individual members of each household: he can choose either to break links which the system would otherwise make or to insert new links.

12. A description of how the contents of the Population Database might be off-loaded into a series of flat files for subsequent loading into a relational database system was provided in Section 11.4.

13. Such a preprocessor has been implemented for converting eighteenth century German passport registration records from their raw input form into a suitably structured form for loading directly into a database management system (Boot et al 1983).

14. I have submitted informal proposals to the Society of Genealogists which are designed to bring about some of the required improvements in this area. See Hawgood 1982, 14–8. The Society subsequently set up an Advisory Panel, concerned with Computers and Public Data, part of whose remit was to establish standards for two distinct data formats. The first, the 'data collection format', has a free format structure and is designed to optimise the data entry process. The second, the 'public data format', has a fixed format structure and is designed to facilitate the distribution and use of the data. This concept of a public data
format was also central to recent technical developments taking place within the 1851 Census National Sample Project (see Note 8).

15. Lee et al 1982. The RESEDA system exhibits a number of interesting features. All data which is to be handled by the system has first to be manually preprocessed. This involves a special encoding operation, in which a human 'analyst' disaggregates the data into its main functional parts (called 'planes') and encodes it using a so-called 'metalanguage'. (In this form it partially resembles the organisation of person occurrence and family occurrence templates, described in Section 10.2.2) Subsequently, the data is loaded into a database, from where it may be accessed by special 'inference procedures'. These procedures, which are intended to mimic the 'intellectual processes of a historian', operate on the data in an attempt to obtain answers to users' ad hoc enquiries.

16. One measure of the complexity of the application is the complexity of the database structure. Thus, the present IDMS schema consists of 3 areas, 49 record types and 46 set types. As such, it represents the most complex database application yet supported by the Edinburgh University Computing Service.

A second measure of complexity is provided by the amount of program code written. The compilation listings for the current system stack to a height of almost 10 inches. Thus, even after making an allowance for the verbosity of the COBOL language this still represents a considerable amount of code.

17. LAMSAC = Local Authorities Management Services and Computer Committee.

18. For an elaboration of my main conclusions see Welford 1977A.

19. The use of such an identifier was discussed in Section 2.1.1.

20. A more complete description of this arrangement and of the required mechanisms is provided in Welford 1977A.
AREA  The total physical storage space used to hold an IDMS database can be logically subdivided into a number of parts, called areas. Broadly speaking, a given area will contain those records and sets which constitute some major subdivision of the database. Our present database subdivides into three areas: the first containing the name directories, the second the source records and the third the population structures.

BOTTOM-UP DESIGN  The characteristic feature of the 'bottom-up' design approach to system development is that one starts by designing and coding a number of primitive routines which carry out tasks which are seen to be necessary in the overall scheme. Progressively the primitive routines are assembled together to create programs of increasing capability and complexity, until ultimately one has a program which can satisfy all the functional demands placed upon it. This approach contrasts with the alternative, 'top-down' approach (q.v.).

CALC  The 'CALC' facility enables efficient access routes into an IDMS database to be established. The method of storing and retrieving each CALC-type record is commonly based on the 'hashing' of some identification field within the record. (For a fuller explanation see Section 4.3.1.)

CODASYL  CODASYL, the Committee on Data Systems Languages, is a U.S. standards body which makes proposals concerning programming language standards (and, specifically, those relating to the COBOL language) and standards for database management system design.

DATABASE  A database is an integrated collection of data which describes or models some entity in the real world, such as a business or a transport network. A characteristic feature of a database is that a given data item, such as a person's address, will normally appear only once in the database. Another significant feature is that the data is arranged in such a way that access to it is convenient and efficient.

DBMS  A DBMS, or Database Management System, is a general-purpose suite of programs which is designed to supervise the organisation of a user's data in a database (q.v.). The DBMS takes responsibility for all the physical aspects of the database, such as how records are
stored within physical files, how they are associated with each other and how they are retrieved. A user of the database is therefore able to ignore all these physical aspects, and by means of a series of special DBMS commands (i.e. the 'DML') he can easily gain access to any required record.

**DDL**

The DDL, or Data Description Language, is the language with which the designer of a particular database initially communicates the structure of the database to the database system. For IDMS the designer must provide details of the areas, records and sets which make up the database. The description of the whole database is contained in the Schema (q.v.), and the Schema DDL is the appropriate language for preparing this. The description of some part of the database which is required by one or more application programs is contained in a Subschema (q.v.), and the Subschema DDL is the corresponding language for preparing this.

**DMCL**

The DMCL, or Device Media Control Language, is the language with which the designer of a particular database communicates the physical storage and input/output buffering arrangements for the database. While the DDL (q.v.) is strictly concerned with the logical organisation of the database, the DMCL is concerned with its physical organisation.

**DML**

The DML, or Data Manipulation Language, is the language with which the writer of a database application program expresses the operations which he wishes to have carried out on the database. To a COBOL programmer, for example, the DML effectively gives him an enlarged programming vocabulary with which to write his programs. Typical IDMS DML database access statements available to the programmer are FIND, STORE, CONNECT and ERASE, and each of these when executed will operate on a particular record in the database.

**DIRECTORY**

Detailed information defining the structure and properties of an IDMS database has to be communicated to the database system using the special DDL and DMCL languages (q.v.). In order to make use of this information the database system creates its own private database, called the Directory, in which it stores the relevant information. This Directory is subsequently referred to when, for example, IDMS is required to process DML statements during the compilation of a program. (This IDMS use of the term 'directory' should be distinguished carefully from its use when describing that part of the record linkage database, viz. the Directory Database, which is used to hold the Christian name, job-name and other name directories.)
FLAT FILE This represents the simplest kind of file organisation. In such a file all the records are of the same size and structure. (See Section 4.2.1.)

HIERARCHICAL FILE This refers to a particular kind of file in which the records are organised in a tree-structured form. Such a file will contain records of different kinds, each having their own size and structure. (See Section 4.2.1.)

IDMS IDMS, or Integrated Database Management System, is a commercially-produced database system, designed according to the CODASYL proposals (q.v.) to handle network data structures.

INTERNAL CODED FORM Nominal records in their original source format have many properties which make them awkward to handle. For example, people's names have different lengths and they are subject to variations in spelling. The process of nominal record linkage can be greatly simplified if the source records are initially 'preprocessed' and converted into a more amenable form. In the present system a program called the Source Translation Subsystem is responsible for this conversion, and the record forms which it produces are described as being in 'internal coded form'. As a result of this conversion each surname, for example, is represented by a fixed-length numeric code value. When information is subsequently to be displayed to the user the internal coded forms have to be converted back into their normal, external equivalents.

MODULE The total system (q.v.) consists of 110 individual sections of COBOL code, called modules. Each module can be compiled separately and is invoked at execution time by means of a COBOL 'CALL' statement. The functional design of the system has been so organised that each module has a limited area of responsibility and a limited number of functions to carry out.

NETWORK A network is a complex structure in which there can be two or more routes connecting any two positions (or 'nodes') within it. A family-tree will, in practice, normally be a network since each individual in it will usually have at least one common ancestor via, say, both his father and his mother. (For an introduction to network database systems see Section 4.2.2.)

RECORD Within an IDMS database the fundamental unit of data storage and retrieval is the record. Each type of record has its own identifying name, which the user employs in his communications with the database system. A record can have a substructure, i.e. it can be made up of several smaller data items, each with its own identifying name.
For example, a PERSON record could contain a NAME data item, a SEX data item, a DATE-OF-BIRTH data item, and so on.

**SCHEMA**

A schema is a description of a whole database, expressed in terms of the areas, records and sets of which it is constituted. It is effectively a master-plan, describing in full the properties of each record type and the potential set relationships which can exist between them. A database designer communicates the schema to the database system using the Schema DDL language.

**SET**

Within an IDMS database the individual records can be associated with each other by means of a set relationship. A particular set type, as defined in the schema (q.v.), will most usually define one particular type of record as the owner record and some other type of record as the member record. A given occurrence of this set in the database will then typically consist of one record of the owner type connected to several records of the member type. Each set type has its own identifying name, which the user employs in his communications with the database system. (For a fuller explanation see Section 4.3.2.)

**SUBSCHEMA**

Individual application programs may not need to be aware of the whole database structure, and so it can be helpful to define a number of simplified 'views', each of which contains only those areas, records and sets needed by the individual programs. Each such simplified 'view' is called a subschema, and the database designer communicates a description of this to the database system using the Subschema DDL language.

**SUBSYSTEM**

Within the system (q.v.) there are six independent programs, called subsystems. Each subsystem operates in isolation from the others, and it has responsibility for some integral part of the total record linkage operation, e.g. source translation. The program code within each subsystem is disaggregated into a number of COBOL program modules (q.v.).

N.B. The Input-Output Subsystem is not a 'true' subsystem, in accordance with this definition. It executes in association with each of the other subsystems, and provides a number of input/output, string-handling and error-reporting functions.

**SYSTEM**

The Record Linkage System consists of all the program code which has been developed within the present research investigation. The total system subdivides into six independent programs, called subsystems (q.v.).

**TOP-DOWN DESIGN**

The characteristic feature of the 'top-down' design approach to system development is that one starts with the
total problem to be solved and then progressively disaggregates it until one is left with the specification of a large number of relatively simple routines which can then be coded. This approach contrasts with the alternative, 'bottom-up' approach (q.v.).
APPENDIX B

PRIMARY SOURCE RECORDS: SAMPLE INPUT FORMATS

B.1 BAPTISMS, 1851-3

SS/PR/BAP/B1/ELWICK HALL PARISH/ANG/30/EXTRACT

B1/12 JAN 1851/WILLIAM/S/JOHN/ELIZABETH/BROOKS/ELWICK VILLAGE/SERVANT
B1/20 APR 1851/MARY ANN/D/THOMPSON/ELEANOR/HALLIMAND/ELWICK VILLAGE/ C
LABOURER
B1/27 APR 1851/OSWALD/S/THOMAS/ELIZABETH/STURDY/BISHOP WEARMOUTH/SAWER
B1/27 APR 1851/ELIZABETH/F/"/"/"/"
B1/2 NOV 1851/Joseph/S/Thomas/Elizabeth/Elder/Amerston Hill/LABOURER
B1/28 DEC 1851/ELIZABETH ANN/D/JOHN/ELIZA/HUTCHINSON/ELWICK MILL/MILLER

B1/14 MAR 1852/MARGARET/ID//MARY ANNE/LEE/ELWICK VILLAGE/SPINSTER
B1/14 MAR 1852/WILLIAM ANTHONY/S/ANTHONY/CATHERINE/THUBRON/BLACKSMITH
B1/18 APR 1852/JOHN/S/JOHN/MARY ROBINSON/ELWICK VILLAGE/ C
PARISH CLERK
B1/6 JUN 1852/HANNAH/D/GEORGE/HANNAH/SUGITT/GUNNER VALE/FARMER
B1/11 JUL 1852/CATHARINE ANN/D/GEORGE FREDERICK/CATHARINE/CHARLTON/ELW C
VILLAGE/SCHOOLMASTER
B1/9 OCT 1853/JOHN WILSON/S/ROBERT/FRANCES/PATTISON/MIDDLE STOTFOLD/ C
FARMER
B1/12 DEC 1851/MARGARET ANN/D/ROBERT/ELIZABETH/ROBSON/AMERSTON HALL/FARMER
B1/27 APR 1851/HANNAH/GEORGE/HANNAH/ALLISON/ENFIELD/TOTFOLD MOORE/HIND
B1/10 JUL 1853/ROBERT/T/ROBERT/ELIZABETH/BROWN/ELWICK/FARMER
B1/24 JUL 1853/MARGARET ANN/D/ROBERT/ELIZABETH FRANCES/CLAYTON/REDGAP/ C
FARMER
B1/23 AUG 1853/WILLIAM/S/THOMAS/MARY ROBINSON/AMERSTONE HALL/FARMER
B1/4 SEP 1853/JOHN ANW/JOHN/RIDDICK/ELWICK/LABOURER
B1/22 AUG 1853/ROBERT TONE/S/ROBERT/HANNAH/ROBSON/BURNTOF/FARMER

B1/17 NOV 1853/ROBERT TONE/S/ROBERT/HANNAH/ROBSON/BURNTOF/FARMER
B.2 MARRIAGES, 1813-22

SS/PR/MAR/MM/ELWICK HALL PARISH/ANG/EXTRACT

M1/D/15 APR 1813/LIC
M1/G/JOHN DOBING/OTP/SIG
M1/B/MARGARET FARTHING/HART PARISH/SIG

M1/D/2 DEC 1813/LIC
M1/G/WILLIAM WALKER/STRANTON PARISH/SIG
M1/B/JUDITH TONE/OTP/SIG

M1/D/18 NOV 1816/BAN
M1/G/THOMAS PICKERING/HART PARISH/BACH/X
M1/B/ISABELLA DOBSON/OTP/SPI/SIG

M1/D/21 SEP 1817/LIC
M1/G/THOMAS SWALWELL/BILLINGHAM/BACH/SIG
M1/B/MARY JOWSEY/OTP/SPI/SIG

M1/D/24 FEB 1820/BAN
M1/G/JOHN WESTCOT?/OTP/BACH/SIG
M1/B/MARY POTTER/OTP/SPI/SIG

M1/D/27 NOV 1820/BAN
M1/G/JOHN HARLE/OTP/BACH/SIG
M1/B/MARY COWANS/OTP/SPI/SIG

M1/D/26 DEC 1820/BAN
M1/G/BENJAMIN DOBSON/OTP/BACH/SIG
M1/B/JANE HASWELL/OTP/SPI/SIG

M1/D/24 FEB 1821/BAN
M1/G/GEORGE FENWICK/HART PARISH/BACH/SIG
M1/B/JANE THUBBRON/OTP/SPI/X

M1/D/10 SEP 1821/BAN
M1/G/WILLIAM RACE/OTP/BACH/SIG
M1/B/ESTHER METCALFE/OTP/SPI/SIG

M1/D/2 DEC 1822/BAN
M1/G/JOHN DAWSON/OTP/BACH/SIG
M1/B/MARY WATSON/OTP/SPI/SIG
B.3 MARRIAGES, 1851-6

SS/PR/MAR/M2/ELWICK HALL PARISH/ANG/EXTRACT

M2/D/17 MAY 1851/BAN
M2/G/JOHN CARR/FA/BACH/LABOURER/GED GAP/X
M2/F/MATTHEW CARR/LABOURER
M2/B/ANN PINKNEY/FA/SPI/SERVANT/BURNTOFT/X
M2/F/THOMAS PINKNEY/LABOURER

M2/D/1 AUG 1853/LIC
M2/G/ROBERT ROBSON/FA/BACH/FARMER/MIDDLE BURNTOFT/SIG
M2/F/GEORGE ROBSON/LABOURING MAN
M2/B/HANNAH STOTHARD/FA/WID/HOUSEKEEPER/MIDDLE BURNTOFT/SIG
M2/F/THOMAS WILSON/LABOURER

M2/D/20 SEP 1853/LIC
M2/G/HENRY ROBINSON/FA/BACH/JOINER/ELWICK/SIG
M2/F/ROBERT ROBINSON/JOINER
M2/B/MARY BRUNTON/FA/SPI/SERVANT/OTP/SIG
M2/F/WILLIAM BRUNTON/LABOURER

M2/D/10 JULY 1855/BAN
M2/G/ROBERT LEE/FA/BACH/SHOEMAKER/ELWICK VILLAGE/SIG
M2/F/WILLIAM LEE/LABOURER
M2/B/JANE SIMPSON/FA/SPI/SERVANT/OTP/SIG
M2/F/ROBERT SIMPSON/LABOURER

M2/D/2 SEP 1855/LIC
M2/G/WILLIAM APPLETON/FA/BACH/LABOURER/BURNTOFT/SIG
M2/F/DAVID APPLETON/LABOURER
M2/B/MAGDALENE ELIZABETH THOMPSON/FA/SPI/FARM SERVANT/BURNTOFT/X
M2/F/JOSEPH THOMPSON/LABOURER

M2/D/15 MAY 1856/BAN
M2/G/JOHN MORROW/FA/BACH/LABOURER/BURNTOFT/SIG
M2/F/NICHOLAS MORROW/FARMER
M2/B/MARY KEARSLEY/FA/SPI/ELWICK/SIG
M2/F/JOHN KEARSLEY/LABOURER
B.4 BURIALS, 1851-5

SS/PR/BUR/D3/ELWICK HALL PARISH/ANG/EXTRACT

D3/ISABELLA CLARK/WOLVISTON/11 JAN 1851/28
D3/THOMAS ALLISON/ELWICK/26 APR/65
D3/DAVID COUSONS/ELWICK/5 MAY 1851/18 W
D3/ANTHONY WILSON/STOTFOLD/6 JULY 1851/26
D3/WILLIAM MIDDLETON/ELWICK/13 JUL 1851/75
D3/ANN ALLISON/ELWICK/27 DEC 1851/5

D3/ELIZABETH CORNER/ELWICK/1 FEB 1852/5
D3/MARY JOSEY/COAL HILL/28 APR 1852/88
D3/MARY ANN SMITH/WOLVISTON/29/20
D3/WILLIAM FRIZZELL/PRESTON/20 MAY 1852/74
D3/ELEANOR STEPHENSON/BILLINGHAM/12 SEP 1852/25
D3/JAMES SMITH/WOLVISTON/10 OCT 1852/19
D3/MARY SMITH/CHURCH SIDE, ELWICK HALL/22 NOV 1852/80
D3/STEPHEN CLEMMET/MIDDLETON HOUSE, HART PARISH/28 DEC 1852/2 M

D3/JANE JACKSON/HIGH STOTFOLD/13 APR 1853/6
D3/MARIANNE SMITH/WOLVISTON/5 SEP 1853/7 M
D3/ANN PREST/ELWICK/18 " "/77
D3/MARY FOSTER/3 CORNERED CLOSE/28 SEP 1853/85
D3/JANE ANN RUDDICK/ELWICK/27 NOV 1853/3 M

D3/JOHN ROBINSON/ELWICK/19 JAN 1854/11 M
D3/HENRY SMITH/WOLVISTON/26 JAN 1854/15
D3/JOHN MIDDLETON/BRUNTOFT/5 FEB 1854/78
D3/THOMAS CORNER/ELWICK LANE/19 MAY 1854/85
D3/JANE BROWN/ELWICK/16 NOV 1854/82
D3/JAMES JOSBON/BENKNOWLE, PARISH OF HART/19 DEC 1854/71

D3/ELIZABETH JANE WELSH/SHERATON/8 MAR 1855/28
D3/MARY SNAITH/THREE CORNERED CLOSE/13 APR 1855/56
D3/WILLIAM ATKINSON/DALTON PERCY/25 APR 1855/3 M
D3/FRANCES MARIA DOBSON/GREATHAM/11 SEP 1855/42
D3/ROBERT DARLING/ELWICK/18 OCT 1855/70
D3/THOMAS HEILMSLEY/HOLE HOUSE/6 NOV 1855/4
D3/WILLIAM ALLISON/ELWICK/11 NOV 1855/75
D3/ANN MARTINDALE/ELWICK/20 DEC 1855/87
SS/CE/C5/ELWICK HALL PARISH

C5/H/1/HIGH STOTFOLD
C5/P/MARY JACKSON/HEAD/WIDOW/68/DURHAM,NEWSON
C5/P/THOS DALE JACKSON/SON/UNMARRD/45/FARMER,ACRES=392,LABS=8/DURHAM, C
/RAMSIDE
C5/P/WILLIAM JACKSON/SON/U/30/FARMER<PARTNER WITH THE ABOVE>/DURHAM, C
/BROOKSIDE
C5/P/ROBERT BELL/SERV/U/20/DO/DO,MARTIN
C5/P/JOHN ELDERS/SERV/U/14/DO/DURHAM,GUNNER VALE

C5/H/2/HIGH STOTFOLD
C5/P/PETER RUTTER/HEAD/MAR/51/FARMER,ACRES=280/DURHAM,SEDGFIELD
C5/P/ISABELLA DO/WIFE/MAR/40/WIFE,OF FARMER/YORKSHIRE,EASINGTON
C5/P/JOHN ATKINSON/SERV/U/26/FARMERS SERV/DURHAM,THORP
C5/P/JOHN HARDY/SERV/U/20/DO/DO,BURDON
C5/P/THO'S NICHOLSON/SERV/U/17/DO/YORKSHIRE,MOULTON

C5/H/3/MIDDLE STOTFOLD
C5/P/ANTHONY WILSON/HEAD/U/26/FARMER,ACRES=250/DURHAM,STOTFOLD
C5/P/AVERAIL DO/SISTER/U/31/DO,DO
C5/P/ANN WALKER/COUSIN/U/24/GENERAL SER V/DO,MIDDLEHAM
C5/P/JAMES HIRD/SERV/U/30/FARM SER V/YORKSHIRE,BELLABY
C5/P/ROBERT FOSTER/SERV/U/14/DO/YORKSHIRE,STOKESLY

C5/H/4/LOW STOTFOLD
C5/P/JOHN LAIDLER/HEAD/MAR/61/FARMER,ACRES=250/DURHAM
C5/P/AZIA DO/WIFE/MAR/69/YORKSHIRE,NK
C5/P/ROBERT PEARSON/SERV/U/14/DO/YORKSHIRE,STOKESLY
C5/P/ISABELLA DO/WIFE/MAR/44/DO,LOW BURNTOFT

C5/H/5/LOW BURNTOFT
C5/P/NICHOLAS MORR/HEAD/MAR/43/HIND,BAILIFF/DURHAM,BILLINGHAM
C5/P/ISABELLA DO/WIFE/MAR/44/DO,BEAMISH
C5/P/WILLIAM DO/SON/U/18/AG LAB/DO,DALTON PERCY
C5/P/ISAAC DO/SON/U/14/DO,DO
C5/P/JOSEPH DO/SON/U/1/DO,LOW BURNTOFT
The illustrations in this appendix serve to demonstrate the results of a particular set of linkage operations. The nominal records which are involved in the linkage operations are listed in Figure C.1. This collection of records consists of all those present in the Source Database which contain a reference to the surname 'ELDERS', or a synonym. The results of the linkage, expressed in terms of some family descriptions, individual life-histories and a nominal index of persons, are listed in Figures C.2 to C.4. A critical examination of these listings and of the underlying linkage operations is provided in Chapter 12.
## 1851 Census Household

<table>
<thead>
<tr>
<th>CO0012</th>
<th>12</th>
<th>Elwick Hall Parish</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Thomas Elders</strong></td>
<td>M</td>
<td>Head</td>
</tr>
<tr>
<td>Elizabeth Elders</td>
<td>F</td>
<td>Wife</td>
</tr>
<tr>
<td>William Elders</td>
<td>M</td>
<td>Son</td>
</tr>
<tr>
<td>Elizabeth Elders</td>
<td>F</td>
<td>Dauph</td>
</tr>
<tr>
<td>Thomas Elders</td>
<td>M</td>
<td>Son</td>
</tr>
<tr>
<td>Robert Elders</td>
<td>M</td>
<td>Son</td>
</tr>
<tr>
<td>Margaret S Elders</td>
<td>F</td>
<td>Dauph</td>
</tr>
</tbody>
</table>

## 1861 Census Household

<table>
<thead>
<tr>
<th>CO0068</th>
<th>35</th>
<th>Elwick Hall Parish</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Thomas Elders</strong></td>
<td>M</td>
<td>Head</td>
</tr>
<tr>
<td>Elizabeth Elders</td>
<td>F</td>
<td>Wife</td>
</tr>
<tr>
<td>Margaret S Elders</td>
<td>F</td>
<td>Dauph</td>
</tr>
<tr>
<td>Joseph Elders</td>
<td>M</td>
<td>Son</td>
</tr>
<tr>
<td>Addison Elders</td>
<td>M</td>
<td>Son</td>
</tr>
<tr>
<td>Sarah Wilson</td>
<td>F</td>
<td>Dauph</td>
</tr>
<tr>
<td>Sister in Law</td>
<td>F</td>
<td>Sister in Law</td>
</tr>
</tbody>
</table>

## 1871 Census Household

<table>
<thead>
<tr>
<th>CO0097</th>
<th>27</th>
<th>Elwick Hall Parish</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Thomas Elders</strong></td>
<td>M</td>
<td>Head</td>
</tr>
<tr>
<td>Hannah Elders</td>
<td>F</td>
<td>Wife</td>
</tr>
<tr>
<td>William Elders</td>
<td>M</td>
<td>Son</td>
</tr>
<tr>
<td>Thomas Elders</td>
<td>M</td>
<td>Son</td>
</tr>
<tr>
<td>John H Elders</td>
<td>M</td>
<td>Son</td>
</tr>
</tbody>
</table>
### Sample Linkage and Analysis Listings

| App. C	 Sample linkage and analysis listings	 621 |
| C00001 |
| (b) |
| ELMACK HALL PARISH |
| 1851 CENSUS HOUSEHOLD |
| P HEAD | M ID | S UN | F ARER | 28 SEP 1824 + 1833 BROOMSTUDE |
| H SERV | F ARER | 27 SEP 1830 + 1833 BROOMSTUDE |
| H SERV | F ARER | 27 SEP 1836 + 1833 BROOMSTUDE |
| H SERV | F ARER | 29 SEP 1832 + 1833 BROOMSTUDE |
| H SERV | F ARER | 27 SEP 1824 + 1833 BROOMSTUDE |

| CO0207 |
| (b) |
| ELMACK HALL PARISH |
| 1862 MARRIAGE |
| * JOHN ELDERS |
| ** THOMAS ELDERS |
| P FATHER | THOMAS ELDERS | 6 OCT 1840 + 1833 ELMACK HALL |
| H SERV | THOMAS ELDERS | 7 OCT 1836 + 1833 ELMACK HALL |
| H SERV | THOMAS ELDERS | 6 OCT 1834 + 1833 ELMACK HALL |
| H SERV | THOMAS ELDERS | 6 OCT 1825 + 1833 ELMACK HALL |

| M00066 |
| (b) |
| ELMACK HALL PARISH |
| 1862 MARRIAGE M2 ANG |
| * JANE PORRIT |
| ** JOHN ELDERS |
| P FATHER | JOHN ELDERS | 13 MAY 1862 LIC |
| H SERV | JOHN ELDERS | 15 JUL 1826 BNS |

---

**Figure C.1** The printing of records for the surname 'ELDERS', by the Record Linkage Subsystem.
<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
<th>Reference</th>
<th>Parish</th>
<th>Date</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1839</td>
<td>Marriage</td>
<td>MO0036</td>
<td>Elwick Hall Parish</td>
<td>2 Dec 1839</td>
<td>BNS</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High Barns</td>
<td>(B)</td>
<td>Bach</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Father: John Brack</td>
<td>(B)</td>
<td>Spinni</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Middleton House</td>
<td>(B)</td>
<td></td>
</tr>
<tr>
<td>1864</td>
<td>Marriage</td>
<td>MO0068</td>
<td>Elwick Hall Parish</td>
<td>20 Aug 1864</td>
<td>LIC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Seaton Carew</td>
<td>(N)</td>
<td>Worr</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Father: Ralph Nuddings</td>
<td>(B)</td>
<td>Spinni</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Amerstone Hill</td>
<td>(B)</td>
<td></td>
</tr>
<tr>
<td>1830</td>
<td>Birth</td>
<td>BO0082</td>
<td>Elwick Hall Parish</td>
<td>(B)</td>
<td>Amerstone Hill</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>M Bap 9 May 1830</td>
<td>9 Apr 1830</td>
<td>30D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Father: William Elders</td>
<td>(B)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mother: Ann</td>
<td>(B)</td>
<td></td>
</tr>
<tr>
<td>1843</td>
<td>Birth</td>
<td>BO0185</td>
<td>Elwick Hall Parish</td>
<td>(B)</td>
<td>Amerstone Hill</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>M Bap 31 Dec 1843</td>
<td>1 Dec 1843</td>
<td>30D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Father: Thomas Elders</td>
<td>(B)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mother: Elizabeth</td>
<td>(B)</td>
<td></td>
</tr>
<tr>
<td>1866</td>
<td>Birth</td>
<td>BO0381</td>
<td>Elwick Hall Parish</td>
<td>(B)</td>
<td>Hart</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>M Bap 22 Jul 1866</td>
<td>22 Jun 1866</td>
<td>30D</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Father: Thomas Elders</td>
<td>(B)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mother: Hannah</td>
<td>(B)</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Event</td>
<td>Reference</td>
<td>Place</td>
<td>Relationship</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
<td>-----------</td>
<td>-------</td>
<td>--------------</td>
<td></td>
</tr>
<tr>
<td>1830</td>
<td>Marriage</td>
<td>M00024</td>
<td>Elwick Hall Parish</td>
<td>John Tinkler * Margaret Elder</td>
<td></td>
</tr>
<tr>
<td>1831</td>
<td>Birth</td>
<td>B00091</td>
<td>Elwick Hall Parish</td>
<td>Ann Elder *</td>
<td></td>
</tr>
<tr>
<td>1836</td>
<td>Birth</td>
<td>B00126</td>
<td>Elwick Hall Parish</td>
<td>John Elder *</td>
<td></td>
</tr>
<tr>
<td>1838</td>
<td>Birth</td>
<td>B00140</td>
<td>Elwick Hall Parish</td>
<td>William Elder *</td>
<td></td>
</tr>
<tr>
<td>1846</td>
<td>Birth</td>
<td>B00210</td>
<td>Elwick Hall Parish</td>
<td>Robert Elder *</td>
<td></td>
</tr>
</tbody>
</table>

Figure C.1 (contd.) The printing of records for the surname 'ELDERS' by the Record Linkage Subsystem
<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
<th>Name of Individual</th>
<th>PARISH</th>
<th>Place of Birth</th>
<th>Town</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1849</td>
<td>BIRTH</td>
<td>MARGARET S ELDER</td>
<td>(B)</td>
<td>AMERSTON HILL</td>
<td>(B)</td>
<td>F BAP 20 MAY 1849 BORN 20 APR 1849 --- 30D FATHER: THOMAS ELDER MOTHER: ELIZABETH</td>
</tr>
<tr>
<td>1851</td>
<td>BIRTH</td>
<td>JOSEPH ELDER</td>
<td>(B)</td>
<td>AMERSTON HILL</td>
<td>(B)</td>
<td>M BAP 2 NOV 1851 BORN 3 OCT 1851 --- 30D FATHER: THOMAS ELDER MOTHER: ELIZABETH</td>
</tr>
<tr>
<td>1854</td>
<td>BIRTH</td>
<td>ADDISON ELDER</td>
<td>(B)</td>
<td>AMERSTON HILL</td>
<td>(B)</td>
<td>M BAP 30 JUL 1854 BORN 30 JUN 1854 --- 30D FATHER: THOMAS ELDER MOTHER: ELIZABETH</td>
</tr>
<tr>
<td>1857</td>
<td>BIRTH</td>
<td>HANNAH ELDER</td>
<td>(B)</td>
<td>AMERSTON HILL</td>
<td>(B)</td>
<td>F BAP 29 NOV 1857 BORN 30 OCT 1857 --- 30D FATHER: THOMAS ELDER MOTHER: ELIZABETH</td>
</tr>
<tr>
<td>1868</td>
<td>BIRTH</td>
<td>THOMAS ELDER</td>
<td>(B)</td>
<td>ELWICK</td>
<td>(B)</td>
<td>M BAP 19 JUL 1868 BORN 19 JUN 1868 --- 30D FATHER: THOMAS ELDER MOTHER: HANNAH</td>
</tr>
</tbody>
</table>
Figure C.1 (contd.) The printing of records for the surname 'ELDERS' by the Record Linkage Subsystem
Sample linkage and analysis listings

ELDERS


JOHN & JANE ELDERS


F00108 JOHN ELDERS


FATHER - P00222 27 SEP 1836 4-1830 GUNNERVALE (B)

MOTHER - P00230 7 OCT 1834 4-1830 UPTON (N)

CHILDREN -

NONE
THOMAS & ELIZABETH ELDERS

FATHER - PO0147  28 SEP 1811 1830 EGGLESTON
MOTHER - PO0148  27 SEP 1812 1830 SEATON CAREW
CHILDREN -
PO0222: JOHN ELDERS M 27 SEP 1836 1830 GUNNERVALE (B)
PO0149: WILLIAM ELDERS M 28 SEP 1837 1830 GUNNERVALE (B)
PO0150: ELIZABETH ELDERS F 27 SEP 1840 1830 WOLVISTON (P)
PO0151: THOMAS ELDERS M 29 SEP 1843 1830 AMHERSTON HILL (B)
PO0152: ROBERT ELDERS M 28 SEP 1846 1830 AMHERSTON HILL (B)
PO0153: MARGARET S ELDERS F 28 SEP 1849 1830 AMHERSTON HILL (B)
PO0154: JOSEPH ELDERS M 7 OCT 1851 1830 ELWICK HALL (B)
PO0155: ADDISON ELDERS M 7 OCT 1854 1830 ELWICK HALL (B)
PO0156: HANNAH ELDERS F 7 OCT 1857 1830 ELWICK HALL (B)

Figure C.2 The printing of details about each conjugal family for the surname 'ELDERS' by the Population Analysis Subsystem
### THOMAS & ANN ELDERS

<table>
<thead>
<tr>
<th>Family ID</th>
<th>Name</th>
<th>Gender</th>
<th>Birth Year</th>
<th>Birth Village</th>
<th>Death Year</th>
<th>Death Village</th>
</tr>
</thead>
<tbody>
<tr>
<td>F00025</td>
<td>THOMAS ELDERS</td>
<td>M</td>
<td>1 Oct 1866</td>
<td>HART</td>
<td>1 Oct 1868</td>
<td>HART</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M</td>
<td>1 Oct 1869</td>
<td>ELWICK HALL</td>
<td>1 Oct 1869</td>
<td>ELWICK HALL</td>
</tr>
</tbody>
</table>

### WILLIAM & ANN ELDERS

<table>
<thead>
<tr>
<th>Family ID</th>
<th>Name</th>
<th>Gender</th>
<th>Birth Year</th>
<th>Birth Village</th>
<th>Death Year</th>
<th>Death Village</th>
</tr>
</thead>
<tbody>
<tr>
<td>F00205</td>
<td>WILLIAM ELDER</td>
<td>M</td>
<td>9 Apr 1830</td>
<td>AMHERSTON HILL</td>
<td>29 Nov 1831</td>
<td>AMHERSTONE HILL</td>
</tr>
</tbody>
</table>

**Figure C.2 (contd.)** The printing of details about each conjugal family for the surname 'ELDERS' by the Population Analysis Subsystem
THOMAS ELDERS

FATHER - NOT KNOWN
MOTHER - NOT KNOWN

MARRIAGES -
FAMILY F00023: WIFE - P00148 ELIZABETH @@@@ @ SONS - 6 DAUGHTERS - 3

EVENTS -
1836 FATHER/BAP: 8 MAY 1836 ANG ELWICK HALL PARISH HIND (B) GUNNER VALE (B)
800126
1838 FATHER/BAP: 22 APR 1838 ANG ELWICK HALL PARISH HIND (B) STOTFOLD (B)
800140
1843 FATHER/BAP: 31 DEC 1843 ANG ELWICK HALL PARISH WOODMAN (B) AMERSTONE HILL (B)
800185
1847 FATHER/BAP: 22 JAN 1847 ANG ELWICK HALL PARISH WOODMAN (B) AMERSTONE HILL (B)
800210
1849 FATHER/BAP: 20 MAY 1849 ANG ELWICK HALL PARISH WOODMAN (B) AMERSTONE HILL (B)
800231
1851 CENSUS: 30 MAR 1851 HEAD MAR WOODMAN (B) AMERSTONE HILL (N)
800231 C000012 12 ELWICK HALL PARISH (B) AMERSTONE HILL (N)
1851 FATHER/BAP: 2 NOV 1851 LABOURER (B) AMERSTONE HILL (N)
800248
1854 FATHER/BAP: 30 JUL 1854 ANG ELWICK HALL PARISH WOODMAN (B) AMERSTONE HILL (B)
800273
1857 FATHER/BAP: 29 NOV 1857 ANG ELWICK HALL PARISH WOODMAN (B) AMERSTONE HILL (B)
800305
1861 CENSUS: 7 APR 1861 HEAD MAR WOODMAN (B) AMERSTONE HILL (N)
800305 C000068 35 ELWICK HALL PARISH (B) AMERSTONE HILL (B)
1868 BURIAL: 30 MAY 1868 ANG ELWICK HALL PARISH (B) STOTFOLD MOOR (B)
800330

Figure C.3  The printing of details about each person called 'THOMAS ELDERS' by the Population Analysis Subsystem.
POO859  THOMAS ELDER

FATHER -  PO0857  WILLIAM ELDER
MOTHER -  PO0858  ANN @@@@@@@

MARRIAGES -  NONE

EVENTS -
1830 BAPTISM:  9 MAY 1830  9 APR 1830 + 30D AMMERSTON HALL

PO0151  THOMAS ELDER

FATHER -  PO0147  THOMAS ELDER
MOTHER -  PO0148  ELIZABETH @@@@@

MARRIAGES -
FAMILY PO0025:
WIFE = PO0158  HANNAH @@@@@@@
SONS = 3 DAUGHTERS = 0

EVENTS -
1843 BAPTISM:  31 DEC 1843  1 DEC 1843 + 30D

1851 CENSUS:
30 MAR 1851  SON SCHOLAR
CO0012  12 ELWICK HALL PARISH

1861 CENSUS:
7 APR 1861  SERV UNM AG LAB
22 JUL 1866  LABOURER
BU0381  1 ANW HALL PARISH

1868 FATHER\BAP:
19 JUL 1868  WOODMAN
BU0399  1 ANW HALL PARISH

1870 FATHER\BAP:
8 MAY 1870  WOODMAN
BU0418  1 ANW HALL PARISH

1871 CENSUS:
2 APR 1871  HEAD MAR FORESTER
CO0097  27 ELWICK HALL PARISH
App. C  Sample linkage and analysis listings  631

Figure C.3 (contd.) The printing of details about each person called 'THOMAS ELDERS' by the Population Analysis Subsystem
ELDERS

ADDISON ELDERS

P00155 ADDISON ELDERS M 7 OCT 1854 ← 183D ELWICK HALL (B)

ANN ELDERS

P00856 ANN ELDER F 29 NOV 1831 ← 30D AMERSTONE HILL (B)
P00156 HANNAH ELDERS F 7 OCT 1857 ← 183D ELWICK HALL (B)
P00158 HANNAH @@@@@@@ F 2 OCT 1846 ← 183D HART (B)
P00858 ANN @@@@@@@ F
<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Sex</th>
<th>Date of Birth</th>
<th>Place of Birth</th>
<th>Place of Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>P00150</td>
<td>ELIZABETH ELDER</td>
<td>F</td>
<td>27 SEP 1840</td>
<td>WOLWISTON</td>
<td>(P)</td>
</tr>
<tr>
<td>P00148</td>
<td>ELIZABETH ELDER</td>
<td>F</td>
<td>27 SEP 1812</td>
<td>SEATON CAREW</td>
<td>(N)</td>
</tr>
<tr>
<td>P00230</td>
<td>JANE PORRIT</td>
<td>F</td>
<td>7 OCT 1834</td>
<td>UPTON</td>
<td>(N)</td>
</tr>
<tr>
<td>P00222</td>
<td>JOHN ELDER</td>
<td>M</td>
<td>27 SEP 1836</td>
<td>GUNNERRYALE</td>
<td>(B)</td>
</tr>
<tr>
<td>P00161</td>
<td>JOHN H ELDER</td>
<td>M</td>
<td>1 OCT 1869</td>
<td>ELWICK HALL</td>
<td>(B)</td>
</tr>
<tr>
<td>P00154</td>
<td>JOSEPH ELDER</td>
<td>M</td>
<td>7 OCT 1851</td>
<td>ELWICK HALL</td>
<td>(B)</td>
</tr>
</tbody>
</table>

Figure C.4 The printing of a nominal index of persons for the surname 'ELDER' by the Population Analysis Subsystem
<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Sex</th>
<th>Date</th>
<th>Age</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>P00557</td>
<td>MARGARET ELDERS</td>
<td>F</td>
<td>1 OCT 1820</td>
<td>72</td>
<td>AMMERSTON HILL</td>
</tr>
<tr>
<td>P00153</td>
<td>MARGARET S ELDERS</td>
<td>F</td>
<td>28 SEP 1849</td>
<td>72</td>
<td>AMMERSTON HILL</td>
</tr>
<tr>
<td>P00656</td>
<td>MARGARET ELDERS</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P00445</td>
<td>MARY ELDERS</td>
<td>F</td>
<td>28 SEP 1771</td>
<td>72</td>
<td>AMMERSTON HILL</td>
</tr>
<tr>
<td>P00152</td>
<td>ROBERT ELDERS</td>
<td>M</td>
<td>28 SEP 1846</td>
<td>72</td>
<td>AMMERSTON HILL</td>
</tr>
<tr>
<td>ID</td>
<td>Name</td>
<td>Gender</td>
<td>Birth Date</td>
<td>Death Date</td>
<td>Location</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>--------</td>
<td>------------</td>
<td>------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>P00147</td>
<td>THOMAS ELDERS</td>
<td>M</td>
<td>28 SEP 1811</td>
<td>183D EGGLESTON</td>
<td>(N)</td>
</tr>
<tr>
<td>P00859</td>
<td>THO'S ELDERS</td>
<td>M</td>
<td>9 APR 1830</td>
<td>30D AMHERSTON HILL</td>
<td>(R)</td>
</tr>
<tr>
<td>P00151</td>
<td>THOMAS ELDERS</td>
<td>M</td>
<td>29 SEP 1843</td>
<td>183D AMHERSTON HILL</td>
<td>(B)</td>
</tr>
<tr>
<td>P00160</td>
<td>THOMAS ELDERS</td>
<td>M</td>
<td>1 OCT 1868</td>
<td>183D HART</td>
<td>(B)</td>
</tr>
</tbody>
</table>

**WILLIAM ELDERS**

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Gender</th>
<th>Birth Date</th>
<th>Death Date</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>P00149</td>
<td>WILLIAM ELDERS</td>
<td>M</td>
<td>28 SEP 1837</td>
<td>183D GUNNERVALE</td>
<td>(B)</td>
</tr>
<tr>
<td>P00159</td>
<td>WILLIAM ELDERS</td>
<td>M</td>
<td>1 OCT 1866</td>
<td>183D HART</td>
<td>(B)</td>
</tr>
<tr>
<td>P00857</td>
<td>WILLIAM ELDERS</td>
<td>M</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure C.4 (contd.) The printing of a nominal index of persons for the surname 'ELDERS' by the Population Analysis Subsystem
PUBLICATIONS AND PRESENTATIONS ARISING FROM THIS PROJECT TO DATE

1977  Project presentation to the IDMS (UK) Users Association meeting in Edinburgh, 17 May.

'Data privacy for everyman: how to avoid the second half of the trip to 1984', Computer Weekly, No 563, 18 August, 18. (This paper won second prize in the 1977 Infotech/Computer Weekly Data Security Competition.)

'The linkage of historical records using the IDMS database system', Proceedings of the Study Group on Computers in Survey Analysis Conference on Data Structures and Software for Survey Analysis, the City University, London, 10-11 November.

1978  Project presentation to International Computers Limited, Dalkeith House, Midlothian, 8 February.

Project presentation, Stockholm Municipal Archive Symposium on Demographic Data Bases, Sweden, 2-3 October.

'Some observations on the future development of the Demographic Data Base Stockholm', Proceedings of the Stockholm Municipal Archive Symposium on Demographic Data Bases, Sweden, 2-3 October, 56-60. (Prepared jointly with Prof M Anderson.)

Project presentation to the Edinburgh Regional Computing Centre, University of Edinburgh, 22 November.

Project presentation to the Student Computing Service, the Open University, Milton Keynes, 4 December.

1979  Project presentation to the SSRC Computing Committee meeting, Abingdon, Oxfordshire, 11 April.


1980  'A project on nominal record linkage by computer', Applied Historical Studies Newsletter, 1980, the Open University.

1981  'Experience of using IDMS for a social science project', SSRC Sponsored Workshop on Database Systems in the Social Sciences, University College, Cardiff, 7-9 July.

1982  'Getting a computer in the family way: the methodological problems of linking nineteenth century censuses and parish
Publications and presentations arising from this project to date

register records by computer', seminar, SSRC Cambridge Group for the History of Population and Social Structure, 15 February.

1983

'The SASPAC project - an assessment, 3: the SASPAC experience', BURISA (Newsletter of the British Urban and Regional Information Systems Association), No 60, September, 10-12. (Prepared jointly with Craig Stott.)

'The establishment of portable interchange formats for genealogical data: can we hope to reach an acceptable standard?', Computer Interest Group of the Society of Genealogists Conference on Computers in Genealogy, Queen Elizabeth College, London, 29 October. (This paper was subsequently published in Computers in Genealogy, Vol 1, No 7, March 1984, 178-87.)

1984
'Standards of data entry', Census Enumerators' Returns in the Age of the Microcomputer, one-day seminar sponsored jointly by the SSRC Cambridge Group and Local Population Studies, Institute of Historical Research, Senate House, London, 21 January.

Establishment of the 1851 Census National Sample as a data library, End of Grant Report, H/00/23/0016, Social Science Research Council. (Prepared jointly with Prof M Anderson.)

'The impact of microcomputers in local history', Scottish Local History, No 4, October, 11-12.

1988
1851 Census National Sample data library: distribution phase, End of Award Report, H/00/23/2032, Economic and Social Research Council. (Prepared jointly with Prof M Anderson.)
LIST OF WORKS CITED IN THE TEXT

I. MANUSCRIPT SOURCES


DCRO (1813) Hart Parish Register. Baptisms 1813-56.


PRO H.O. 107/2384. Includes enumerator's schedules for Elwick Hall and Hart parishes (1851).

PRO R.G./9 3699. Includes enumerator's schedules for Elwick Hall and Hart parishes (1861).

PRO R.G./10 4913. Includes enumerator's schedules for Elwick Hall and Hart parishes (1871).

PRO R.G./10 4918, 4919 and 4921. Enumerators' schedules for part of Stranton Parish, West Hartlepool (1871).

II. MISCELLANEOUS LOCAL PRINTED SOURCES


III. MACHINE-READABLE SOURCES

Each of the following sources is drawn from the 1851 Census National Sample for Great Britain, PDF1 format. The machine-readable files are available from the ESRC Data Archive, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ.

PRO H.O. 107/1691. Includes enumerator's schedules for the parish of Bucklebury, Berkshire. Cluster number 1025.
IV. PARLIAMENTARY PAPERS

1852-3 LXXXVI Census of England and Wales, 1851. Population tables, part 1. Number of inhabitants in the years 1801, 1811, 1821, 1831, 1841 and 1851.


V. OFFICIAL PUBLICATIONS


VI. BOOKS, MANUALS, ARTICLES, ETC.


ANDERSON, M (1972) 'The study of family structure', in WRIGLEY (1972), 47-81.


BLAYO, Yves (1973) 'Name variations in a village in Brie, 1750-1860', in Wrigley (1973), 57-63.


GINTER, Donald E, GROGONO, Peter and BODE, Frederick A (1977) 'A review of optimal input methods: fixed field, free field, and the edited text', Historical Methods Newsletter, Vol 10, No 4, Fall 1977, 166-76.


HERLIHY, David (1973) 'Problems of record linkage in Tuscan fiscal records of the fifteenth century', in Wrigley (1973), 41-56.


SKOLNICK, Mark (1973) 'The resolution of ambiguities in record linkage' in WRIGLEY (1973), 102-27.


TILLOTT, P M (1972) 'Sources of inaccuracy in the 1851 and 1861 censuses', in Wrigley (1972), 82-133.


WINCHESTER, Ian (1973) 'A brief survey of the algorithmic, mathematical and philosophical literature relevant to historical record linkage' in WRIGLEY (1973), 128-50.

List of works cited in the text


WRIGLEY, E A (ed) (1973) Identifying people in the past, Edward Arnold.

