NATIONAL SAMPLE FROM THE 1851 CENSUS

OF GREAT BRITAIN

INTRODUCTORY USER GUIDE

Michael Anderson

Pre-release edition
September 1987

Department of Economic and Social History
University of Edinburgh
William Robertson Building
George Square
Edinburgh
EH8 9JY

INDEX

PREFACE

This document provides a general introduction to the National Sample from the 1851 Census of Great Britain project. More detailed material can be found in the documentation listed in Section 10.

I had the idea for this project one summer evening in 1971. Since then an enormous number of people have allowed themselves to become trapped into contributing to it in one way or another. Brenda Collins and Craig Stott bore the brunt of the traumas of making the idea into a lusty (400,000 person) but infant reality. John Welford provided the inspiration which allowed the infant to grow up and Linda Croxford and Linda Aitken put some preliminary clothes onto John's ideas. Barbara Petrie and Alison Morrow have finally achieved what for years seemed the impossible mature and final product, the unglamorously named but enormously serviceable Public Data Format 2 National Subsample.

Along the way we have all built up an enormous number of debts to the others who have helped us. Ros Davies of the Cambridge Group provided the software which got us going. Susan Robson toiled away at the excruciatingly boring task of drawing the sample in the Public Record Office. More than a dozen ERCC punch operators gave up much of the best years of their lives in turning illegible xeroxes into machine-readable transcripts; Bill Gordon and Kate Niven bore the brunt of their anxieties and ensured that an excellent job was done. Craig Dickson, Barbara Morris, David Munro, Susan Hutchison, Ruth Brown and Debbie Kemmer produced between them most of the PDF1 files and the associated documentation sheets and directories. The Steering Committee for the third grant (Clive Payne, Richard Wall, Richard Bland and Eric Roughley) gave up freely of their time to advise and encourage us and have played a vital role in shaping the final product; Eric Roughley had known us and our mass of files for much longer, and has still remained cheerful to the end.

We should also acknowledge the generous and understanding support of the SSRC/ESRC which has funded us intermittently over a period of fifteen years. If they had accepted the original proposal for a two year grant for under £9000 they would have saved everyone a lot of effort, but the end result would have been trivial compared with what we have achieved. The Edinburgh Computing Service has supported us most generously, above all by meeting (if at times just a shade disbelievingly) our incessant requests for yet more file space. Finally, the Departments of Sociology and of Economic and Social History have been willing and supportive hosts to the projects.

Michael Anderson

1. THE 1851 CENSUS OF GREAT BRITAIN AND ITS PROCEDURES

The 1851 Census of Great Britain was conducted on Sunday 30th March 1851. In the preceding days almost seven million schedules had been distributed to 'occupiers' of residential property by 38,483 enumerators. Schedules had also been distributed to military installations and to public institutions of all kinds.

The census report records (though this was probably somewhat optimistic) that enumerators experienced few difficulties in the issuing or collection of the schedules. However, the general impression that is left from an extensive study of the records over many years is that almost all the enumerators did at least as good a job as could have been expected given the conditions under which they had to work. The schedules were either filled in by the occupier on behalf of his or her household or, in cases of illiteracy or lack of cooperation, were completed by the enumerator in response to questions to one or more residents.

The schedule was a marked improvement on that used in 1841, the first British census taken by modern enumeration techniques. It allowed more space for answers and sought much more precise information. In brief, each occupier was required to provide, for each member of the household, information on: prename, surname, relationship to head, marital status, sex, age, occupation, birthplace (by parish and county if born in the country of enumeration), and on infirmities of speech, hearing, and/or sight. In addition, information was collected on addresses, and enumerators were instructed to record the presence of properties that were empty or under construction, and to distinguish what we today would call secondary households within houses. An attempt was also made to count the numbers of people who had spent census night in places other than houses and institutions ('in barges, boats or other small vessels remaining stationary on canals and other navigable waters', 'in barns, sheds or the like', 'in tents, or in the open air').

Once they had collected the schedules, enumerators were instructed to number them and then to transcribe them into their 'enumerators' books'. These books were then passed, together with the schedules, to the local official responsible for the census who checked them for completeness and, where possible, accuracy. At this stage a number of modifications were made which are clearly visible on the original documents. Further checking, and sometimes further annotations, were then added by senior local officials. The books and schedules were then forwarded to London where at some point the original schedules were

destroyed. A small army of census clerks (sometimes referred to as 'checkers') was employed to extract manually the information from the enumerators' books and to make abstracts onto large blank tables from which the final published totals were prepared. The checkers also marked the books, indicating the extractions that had been made. They sometimes also added brief comments indicating how they had handled ambiguous or incomplete responses. Some of these annotations - which mainly involve putting places into counties where no county was specified, and allocating imprecisely defined occupations to industry groups - can in the light of local knowledge be revealed to be nonsensical; most of their judgements are, however, defensible if one remembers that it was the overall national or local picture rather than the situation of any individual family that was their main concern, and that in this context most errors probably cancel out.

In one particular aspect of the checkers' work extreme arbitrariness was the only available solution. The table in which enumerators recorded the numbers of people living outside normal residential property (Table b in the enumerators' books) asked enumerators only to divide the population by sex. The checkers' tables for 'Age', however, had no category for 'age unknown or not recorded' and this meant that, wherever age information was unavailable, ages had to be guessed at. This was relatively rare in normal households but for the Table b entries a formula was needed. This involved allocating ages in a cycle, usually in the following order: 25, 35, 15, 45, 5, 55, 65. Thus if sixteen individuals were present, three each would be allocated to ages 25 and 35, and two to each of the other age headings.

2. THE 1851 CENSUS NATIONAL SAMPLE PROJECT

The nineteenth century census enumerators' books have been extensively used for many years both by genealogists and by scholars working on studies of local communities. In the 1960's the scale of these local studies was significantly enlarged (and studies of substantial towns made possible) by the employment of sampling techniques to extract representative subsets from the data on any place, and by the use of electronic methods - card sorters and then computers - to process the results. The findings of these local census-based studies, both in Britain and elsewhere, were of enormous significance and played a major role in the 1960's and 1970's in the reorientation that occurred in both the techniques of social and economic history, and in its concerns and themes of study.

Locally-based study had, however, two drawbacks, which limited the full impact that the census enumerators' books could make on our knowledge of the past. The first limitation arose because, in the absence of a national picture against which to set the conclusions of local work, it was often difficult to appreciate how significant any particular set of findings might be. In addition, where a national picture was required (for comparison, for example, with modern studies) this could only be approximated by some sort of averaging process across the findings of the various locally based pieces of work. Secondly, where the focus of interest was not primarily local but was concerned with a particular section of the population (shopkeepers or professional men, for example), no clear picture could be produced by locally based work, but the task of conducting a wide-ranging national sample just for one small section of the population was totally ruled out on cost grounds.

It was against this context that work began in Edinburgh in 1972, with support from the then Social Science Research Council, to create a machine-readable national sample from the enumerators' books of the census of the whole of Great Britain. A machine-readable transcript was to be prepared and from this a range of data sets would be created which would allow the data to be processed rapidly, using standard data analysis software, by scholars with a minimum of technical expertise, and working on their local computer systems. In the event this task took fifteen years, with the last of the major data sets being deposited in the national ESRC Data Archive at the University of Essex in 1988; the Data Archive now provides the main source of information on access and contents of all the large scale computer-readable data sets, and inquiries such data should be directed to it. As is outlined in the following sections, the census data

exist in a number of different forms, ranging from the original machine-readable transcripts, to a hierarchically structured 'Public Data Format' which, when used with associated postprocessor software, can produce data which can be passed direct to commercial database management systems.

Meanwhile, in the intervening years, data have been distributed in a number of other ways. In particular, through the Edinburgh University Data Library Service, small scale teaching data sets are provided in QUEST-compatible format for use on BBC micros in schools and colleges, and similar data sets are being developed for use as component parts of undergraduate courses in the University of Edinburgh, with the intention of exporting them to other institutions. At the initiative of the Open University, printouts of data on particular places, arranged in a tabulated format closely matching the original enumerators' book layout, are provided for student project work and this service is also available to other educational institutions. A microfiche reproduction of the data in this layout, published by Chadwyck-Healey, will also make available to the general user the entire National Sample data set in this tabulated format.

3. THE SAMPLES

Sampling procedures for the main sample

The main sample is a two percent stratified systematic cluster sample. This means that in order to maximise the extent to which the sample was representative of different kinds of places, all settlements listed in the published census reports were first stratified (divided) into groups; these consisted of 'towns', 'small non-urban settlements', 'large non-urban settlements', and a residual category of 'other places'. Institutions listed separately in the published Reports formed a fifth group (or stratum). The first category, 'towns', consisted of all the settlements located within the boundaries of Municipal Boroughs and of other places identified as 'towns' by the census authorities. Settlements within Parliamentary Boroughs were also included in the 'urban' category, though in a few cases this led to inclusion of some areas which might more normally be considered as non-urban. The second category, 'small non-urban settlements', consisted of all settlements separately listed in the census Reports with a population in 1851 of 2,000 or less. All remaining complete settlements were categorised as 'large non-urban', except for a few areas within settlements which were outside town or borough boundaries; these make up the residual 'other places' category.

The main sample was, with two major exceptions, drawn by selecting every fiftieth enumeration book which related to each category of place. The first exception was that, in England and Wales only, the small non-urban settlements were sampled by taking the whole of every fiftieth appropriate place listed in the published census; this part of the sample thus provides in its own right a sample of smaller English and Welsh villages and can be used in a number of interesting ways to throw light on the nature of the society of this very characteristic English/Welsh phenomenon. Secondly, institutions were sampled by treating all the institutional populations as if they were one continuous list and then systematically selecting twenty individuals from each successive one thousand names. Where institutions included families, special arrangements were made; in these cases the sample was drawn so as to include all members of families the first member of which appeared before the twentieth individual, and to exclude members of families where only the later members appeared within a block of twenty names.

Proceeding in these ways, 980 separate 'data clusters' were identified, either directly from the published Reports or, in most cases, by

tediously counting the enumeration books and selecting the fiftieth in each stratum sequence. Photocopies of all the material required were then obtained from the Public Record Office in London or from New Register House in Edinburgh, except in a very small number of cases in which the records had not survived or were not in a fit condition to be transcribed. In all, over 30,000 separate photocopied sheets were eventually obtained and prepared for punching into machine-readable form. The data from these 30,000 enumerators' book pages make up the 'National Sample from the 1851 Census of Great Britain'

A warning on sampling error

This data set was generated to be a nationally representative sample. At the wish of SSRC it was also designed as a cluster sample. This raises significant problems of inference even for national studies (since the computation of sampling error from large, heterogeneous, variable sized-sampling unit, cluster samples is a complex business. It raises even more problems for anyone wishing to undertake a local or regional study.

The first point to be remembered is that these data are in general a sample of enumeration districts only. It should therefore be obvious that the sampled book or books for any particular town or village cannot validly be used, except in a very exploratory or rough manner, to make inferences about the situation of the whole of that town or village's population in 1851. One exception here is the case of the smaller non-urban settlements for which the complete data have been collected in each cluster; another is London, since exploratory research does suggest that its 51 books do provide both a large enough and representative enough sample to allow reasonable inferences to be made about London's population as a whole. Less obviously, but equally true in many cases, the number of clusters available is also too small to allow valid statistical conclusions to be drawn about the populations of most counties (Lancashire and Yorkshire should be exceptions, though the vagaries of systematic sampling have in fact made the Yorkshire data somewhat less representative than one might have liked). As a rough guideline, experience suggests that if one wishes to ensure even a roughly representative set of data one probably needs to draw material from at least thirty data clusters.

Users should therefore note that drawing inference from any of these data (including and particularly the subsample data described below) is a non-trivial problem. If any doubts at all are felt by a user about the validity of inferences being made, then professional statistical advice should be sought.

Main Sample cluster numbers

Each main sample data cluster was given its own unique identifier. Since these are the numbers which are used to identify the main sample data files whenever they are accessed, some understanding of their significance will be useful to at least some users.

The entire data collection is first grouped into the twelve regional divisions used by the census authorities; all files relating to each division fall into numerical bands as follows:

0601-0999 London
1001-1499 South Eastern Counties (Surrey, Kent, Sussex, Hants, Middlesex)
1501-1999 South Midland Counties (Middlesex, Herts, Bucks, Oxford, Northants, Huntingdon, Bedford, Cambridge)
2001-2499 Eastern Counties (Essex, Suffolk, Norfolk)
2501-2999 South Western Counties (Wilts, Dorset, Devon, Cornwall, Somerset)
3001-3499 West Midland Counties (Gloucester, Hereford, Shropshire, Stafford, Worcester, Warwick)
3501-3999 North Midland Counties (Leicester, Rutland, Lincoln, Nottingham, Derby)
4001-4499 North Western Counties (Cheshire, Lancashire)
4501-4999 Yorkshire
5001-5999 Northern Counties (Durham, Northumberland, Cumberland, Westmorland)
6001-6499 Wales
7001-7499 Scotland

Within each of these divisions the last three numbers of each cluster code are allocated according to the relevant stratifying criterion:

001-099 and 501-599 Small non-urban places
101-199 and 601-699 Towns (sequences from x151 and x651 relate to clusters from Parliamentary Boroughs)
201-299 and 701-799 Large non-urban places
301-399 and 801-899 Other places
401-499 and 901-999 Institutions

Within each of these blocks the clusters are numbered sequentially in the order in which their settlements were listed in the published reports (and in the county order within divisions as listed above). Thus, for

example, clusters 1001, 1508 and 3016 are all 'small non-urban places', located respectively in Surrey, Oxfordshire, and Shropshire; since they are all in England, they are sampled as settlements in their entirety. 7015 is a single enumeration district from a small non-urban place in Scotland. Cluster 4151 is an enumeration district from a Parliamentary Borough of Bury in Lancashire. Cluster 0909 comprises twenty individuals from the Westminster House of Correction in London, and so on. The few gaps in some of the sequences relate either to data which turned out not to be accessible or in a fit state to be transcribed, or to clusters which have subsequently been reallocated to other strata following closer inspection of their boundaries and populations.

The Subsamples

For many purposes the entire sample is too large to be handled conveniently as a data set. Lack of resources also made it clear from an early stage that it would not be possible to process fully all the data in all the ways that we should have liked.

Our first approach to subsampling was to process completely all the data from a proportion of the clusters. We thus developed a test data set which contained 65 clusters, the subsampling being performed initially by taking every sixteenth cluster. Investigation of the properties of the resulting data set revealed that there were major problems of representativeness in the manufacturing districts. These were solved by boosting which invloved taking every eighth cluster in the manufacturing districts and then weighting the results for all manufacturing district clusters at one half in the course of analysis. In all, about 8% of the total sample was included in this first subsample. A great deal of experience was gained from working with this data set, some tentative results from it were published, and a version of it was deposited in the Data Archive.

However, further consideration of the subsampling issue with our statistical advisers led us subsequently to adopt a rather different approach since the whole cluster method produced major problems when it came to estimating sampling error. As a result, we adopted a strategy which involved drawing a series of subsamples of households from the total data set by means of replicated sampling. A computer program was written which allowed us to extract eight systematic 1 in 40 samples from the entire data set; each of the subsample replicates contained about 10,000 individuals. These samples have been stored as a series of discrete files, with one file for each replicate for each stratification stratum for each division (except that institutional data were grouped into two broad north-south groups for the non-London areas of England and

Wales). For example, the first replicate sample of one in forty of all the population of the London books is in a file called SSA06, the second replicate in SSB06, the third in SSC06. All the urban clusters for Division 2 are in a series of files called SSA11 through SSH11, all the Scottish institutions in a series of files called SSA74 through SSH74. All these basic subsample files are stored in a specially structured binary form on the Edinburgh FILESTORE.

A proportion of these files has then been further processed into the Public Data Format described in a later section. Three of the eight complete replicates of the small non-urban places clusters, and six complete replicates of all the other data, have been processed in this way. Taken together these files form the 'National Subsample Files' and are likely to be the most valuable and important data sets for most users in the future. The numbering systems applicable to these files are derived from those used to described the original subsample replicates and are explained in a later section below.

4. PREPARATION OF THE MACHINE-READABLE DATA-SET

Data preparation procedures

The preparation of the machine-readable data set was a lengthy and tedious task which occupied initially some four years between 1973 and 1977. The object specified by the SSRC was initially to prepare a machine-readable transcript which conformed as closely as possible to the format, content and structure of the original enumerators' books. The punch operators were therefore instructed to try to punch exactly what they read in the documents, not to seek to interpret or translate what they saw, and above all to make no attempts to 'correct' or standardise the text.

As a result, for several reasons, the resultant data files are not 'accurate' representations of the social structure of the communities surveyed. Firstly, some of the original respondents were clearly ignorant of their precise ages or birthplaces; they also no doubt at times felt it appropriate to give answers which were to varying degrees 'false' (that they were married when they were in fact cohabiting, that their eight year old children were not employed in textile mills, and so on). Sometimes the ways in which the questions were posed discouraged 'accurate' responses (for example women's involvement in casual or part-time occupations is clearly under-recorded for this reason). The illiterate gave spoken responses to the enumerators and seldom knew the correct spelling of their names and birthplaces; it is thus not surprising that enumerators frequently gave recognisably phonetic versions of placenames and sometimes wrote things which are quite uninterpretable at least to those who undertook the transcriptions. Sometimes the enumerators could not read exactly what was written on the schedules that had been handed to them and what we have is their best guesses. Frequently, we believe, when they transcribed the schedules into the enumerators' books, they reordered the houses to fit exactly on the pages of the books and they also changed in various ways what had been written on the schedule to make the style of the book consistent internally or more in line with what they understood as the instructions that they had been given; sometimes (as with the division of properties into houses and of houses into households) these instructions were themselves vague or unworkable. Occasionally, the enumerators made transcription errors themselves.

Finally, and inevitably, 'mistakes' were made in the creation of our machine-readable transcripts, in spite of considerable attempts to ensure that they were as accurate as possible. Before punching, the photocopy of

each book was read over by one of the research staff and attempts were made to foresee likely areas of difficulty either in the hand-writing or in the interpretation of what was written; where problems were anticipated, the photocopy was marked with the version that was to be punched. To assist in this process a number of gazetteers and other reference books were available, but these were used as guides to likely interpretations rather than to correct what had been written even in cases where the enumerators had obviously 'got it wrong'. The punch operators then punched what they thought they read, though again they were free to consult with the research staff where they had problems in interpretation. After punching, and depending on the difficulty of the book and the experience of the operator, the data were then either subjected to a physical check by repunching so as to 'verify' the data, or the output was inspected visually on the terminal by the operator before being written to the file. Next, the data were run through an earlier version of the listing program which has since been used to generate the microfiche and the Open University student printout, and the printouts were visually read over to check for glaring inconsistencies and omissions. At a later stage the format was checked by specially written software; many other errors were identified and corrected during the processing necessary to generate the coded files which were developed for large scale computerised analysis. Most recently, a series of validation routines and manual checks have been run over the data which comprise the National Subsample Files. One consequence of this is that the latest versions of the data will not always correspond exactly with some earlier publically available data sets.

While, in spite of all these checks, some transcription errors remain (and we are continuing to update and improve the quality of the data files from year to year), experience now suggests that most of the 'errors' which are being identified by users of the data are in fact either obviously present in the original data or are arguable alternative readings of poor quality text. We nevertheless encourage users of the machine-readable files and of our printouts and microfiche to send us 'corrections' and we hope that all users will continue to do so.

Punching conventions for the transcript data

A copy of an original page of an enumerators book appears as Figure 1 and a copy of part of the transcript of this page as Figure 2. As indicated above, the first object of the project was to produce a transcript which conformed as closely as possible to the original enumerators' books. The transcript was thus prepared with the spelling and abbreviations as far as possible unchanged from the source, and, for example, with 'dittoes (as

long as they were 'Do', 'Do Do', 'Do Do Do', ' Ditto', '"', '"" ""', or '"" "" ""') included wherever they appeared in the original. In certain cases, however, a number of conventions had to be adopted to deal with inconsistencies in the original data, to handle various visual features which could not be readily converted into text, and to maintain in the files the internal consistency necessary for straightforward computer processing. Some understanding of these alterations is helpful for users of the data in no matter what form they receive it.

As originally punched, the data were intended for input into a set of simple hierarchical data-handling routines (GENDATA) which had originally been developed by the Cambridge Group for processing parish register data on IBM 360 series machines. GENDATA operated on a chain and element data structure, where each chain type was indicated by a single character in the first field of a card-image, and where the elements were then punched in free format with the symbol '/' being used as the delimiter (see below); there was also a facility which allowed for optional elements to be present in only some parts of some records, and these optional fields are delimited in our data by the symbols '<' and '>'. Data were punched in the first 72 'columns' of each card-image, but with the possibility of overflow onto subsequent 'cards'. Chains and elements were grouped into 'records'; in our data the household is the record unit and was delimited from the next record by the insertion of a '$' symbol in the first column of an otherwise blank card-image. Information on the household as a whole was stored in elements on the household chain (indicated by the initial character 'Z'). Information on each individual within the household was stored on the person chain (indicated by the initial character 'J'). A household which appears in the original enumerator's book as follows:

48	18 Earl St	William	Brooks	Head	Mar	59		Dresser of warps	Lancashire	Bury
		Nancy	Do	Wife	Mar		60		Do	Bury
		Adam	Do	Son	U	38		Weaver (Cotton)	Do	Haslingden
		Ann	Do	Daur	U		19	Weaver (Cotton)	Do	Bury
		William	Do	Grandson	U	8		Scholar	Do	Do
		David	Kirkman	Lodger	Mar	35		Labourer at foundry	Do	Do

was punched as follows:

```
$
Z/11/48/EARL ST/8
J/WILLIAM/BROOKS/HEAD/MAR/59/-/DRESSER OF WARPS/LANCASHIRE/BURY
J/NANCY/DO/WIFE/MAR/-/60/-/DO/BURY
J/ADAM/DO/SON/U/38/-/WEAVER (COTTON)/DO/HASLINGDEN
J/ANN/DO/DAUR/U/-/19/WEAVER (COTTON)/DO/BURY
J/WILLIAM/DO/GRANDSON/U/8/-/SCHOLAR/DO/DO
J/DAVID/KIRKMAN/LODGER/MAR/35/-/LABOURER AT FOUNDRY/DO/DO
```

This is a fairly straightforward example, but during the punching

operation a whole series of conventions had to be employed to deal with various aspects of the data. Each of these is now briefly discussed in turn.

a) Illegibility:

Wherever the text inserted seemed uncertain the punched field ended with a '?' (e.g. DEUTSCHKREUTZ?); where some letters were also unclear question marks were substituted for the doubtful characters (e.g. DEUTS??KREUTZ?); where whole words were illegible and no reasonable guess could be made then '??' was inserted into the field (a few books, much damaged by damp in storage, produced photocopies so faint that much of their content could only be transcribed in this way).

b) Coding non-transcribable information:

In various situations it was necessary to insert 'coded' numbers or other symbols to modify the original data because, although the intention of the original was clear, a machine-readable textual transcription was either impossible or would not produce meaningful results when read by the computer on an individual-by-individual basis. The main modifications were:

i) Coding the lines or other indicators used by enumerators to distinguish houses and households. The precise procedures used here are described in a later section of this document.

ii) The elaboration of certain entries in the text where this was necessary if an individual's record was to be comprehensible when removed from its visual context. While, as noted above, the computer was programmed to handle dittoes, from time to time the original text contained empty fields which were then dittoed, or cases where only part of an entry was dittoed. For example, an original reading:

John McDougall	Servant	Married	42	Farm Servant	Ayrshire	Ayr
Jane Mc Do	Do wife	Do	40		Do	Do
James Mc Do	Do son	Un Do	18	Do	Do	Do

had to be punched as follows:

```
JOHN/MCDOUGALL/SERVANT/MARRIED/42/-/FARM  SERVANT/AYRSHIRE/AYR
JANE/DO<MC DO>/SERVANTS WIFE<DO WIFE>/DO/-/40/-/DO/DO
JAMES/DO<MC DO>/SERVANTS SON<DO SON>/UNMARRIED<UN DO>/18/-/FARM SERVANT<D%>/DO/DO
```

There are several problems in the original format which have been solved in the transcription. Entries like 'Do wife' are contextually specific (if John had been a lodger then 'Do wife' would mean 'Lodger's wife') 'Mc Do' is equally unclear out of context. In each case the entry has therefore been spelled out in full. James's occupation is also unclear out of context; indeed strictly speaking, like his mother, he has no occupation and the computer would certainly seek to translate the field as such. Again,

therefore, a full entry has been transcribed. Note, however, that in all such instances the original entry is spelled out in angled brackets (<.....>) in a comment field. The entry '<D%>' replaces '<DO>' because the software was not designed to cope with dittoing empty fields; '*' replaced '"' under similar circumstances.

iii) There were also occasions where, for example, the enumerator had bracketed together entries relating to two or more people (e.g. with the comment 'twins') or where he had inserted asterisks to indicate omissions. These situations were translated as appropriate into what it was hoped was unambiguous text; frequently an entry was also made in the comment field for the household (see below).

c) Standardising the order of fields in the text:

In order to maintain consistency for computer processing it was necessary that the fields appeared in the data set in a standard order; this meant that blank fields (indicated by hyphens) had frequently to be inserted in cases where enumerators had left some of the columns empty or had omitted, for example, to give a surname. Problems also arose in the following cases:

i) With birthplaces where it was necessary to maintain the standard order of the enumerators' instructions (i.e. County/Parish/Country).

Thus:

'Ireland'	was punched as: -/-/IRELAND
'Lerwick Zetland'	was punched as: ZETLAND/LERWICK<21>
'Zetland Scotland'	was punched as: ZETLAND/-/SCOTLAND
'Scotland Zetland'	was punched as: ZETLAND/-/SCOTLAND<31>
'Ayr Ayrshire Scotland'	was punched as: AYRSHIRE/AYR/SCOTLAND<213>

The numbers in the angle bracket comment field indicate in each case the order of the entries in the original text.

ii) Similar adjustments were at times required to maintain a consistent order of prenames and surnames (prename always precedes surnames in the punched data), and of street names and house names (or numbers) or area of a settlement and its specific address within the place (e.g. it was necessary to enter 'GEORGE STREET/18', instead of '18, George Street' and 'GRUTTING/MILL HOUSE', instead of 'Mill House, Grutting').

d) Insertions to maintain data consistency:

In a limited number of cases, in order to facilitate computer processing some information was substituted for the original text, which was then bracketed in angle brackets. The main examples of this are:

i) Where age entries contained non-numeric entries (e.g. 'one month'), the information was punched in numeric (decimal where relevant) form. 'Not known' was punched as '999' and 'infant' as 998. In all cases the original is bracketed in the transcript.

ii) In the birthplace columns, even where the enumerator had not done so, a county of birth was included in the data set wherever this could be done unambiguously. A special problem arose where place names were the same as county names (e.g. Northampton, or Dumfries) and where the enumerator had only inserted the name once, usually in the middle of the column. This was handled by assuming in these cases that the person was born in the town in the county, a solution which involves some errors; however, for later processing reasons some decision had to be made and this seemed the least misleading assumption to make. In all cases insertions are bracketed in the transcript.

iii) In a few cases, and regardless of the general principles of producing as far as possible a literal transcript, it was necessary to 'correct' the enumerators' entries. Thus, for example, where they were detected during preparatory processing, relationships to head of household were changed to prevent a head having two wives where it was in fact obvious that the second wife was the wife of a lodger or of a relative in the household. In conformity with the rules developed by historians dealing with census data, 'lodgers' were not allowed to be heads of households, and appropriate adjustments were made to maintain consistency. Where sex conflicted with relationship (e.g. female sons) consistency was restored using name or other information.

e) Registrars' and checkers' marks:

As was indicated in the opening section, the enumerators' books also contain a number of marks and comments in other hands. Where these were legible they have been inserted into the data files. Where they relate to whole individuals or households or to the book as a whole they have been transcribed into the special household comment fields (see below). Where they relate to an individual field, the census office checkers' marks appear in brackets in the transcript and are not incorporated into the text, but where local superintendents and registrars have made amendments of a

kind clearly based on local knowledge and, for example, elaborating a cryptic enumerator's entry, they have been incorporated in the basic text and the enumerator's original has been bracketed. In one case two slightly discrepant transcriptions of the same set of schedules have survived. In this case, where entries in the duplicate differ from those in the book forwarded to the census office they are bracketed in the transcript and indicated with the special symbol '*'.

f) Reserved or unprintable characters:

A small number of further amendments was required because the text character involved was not part of the standard computer character set or was required for some special purpose by the data structuring or coding programs.

Thus:

'$^1/_2$' was punched as '.5'
'/' was punched as '(' [or as a space]
'-' was punched as a space
''' was ignored though subsequently new software allowed the reinsertion of some entries
'.' was ignored or punched as ':'
',' was ignored or punched as ';'

The file documentation data

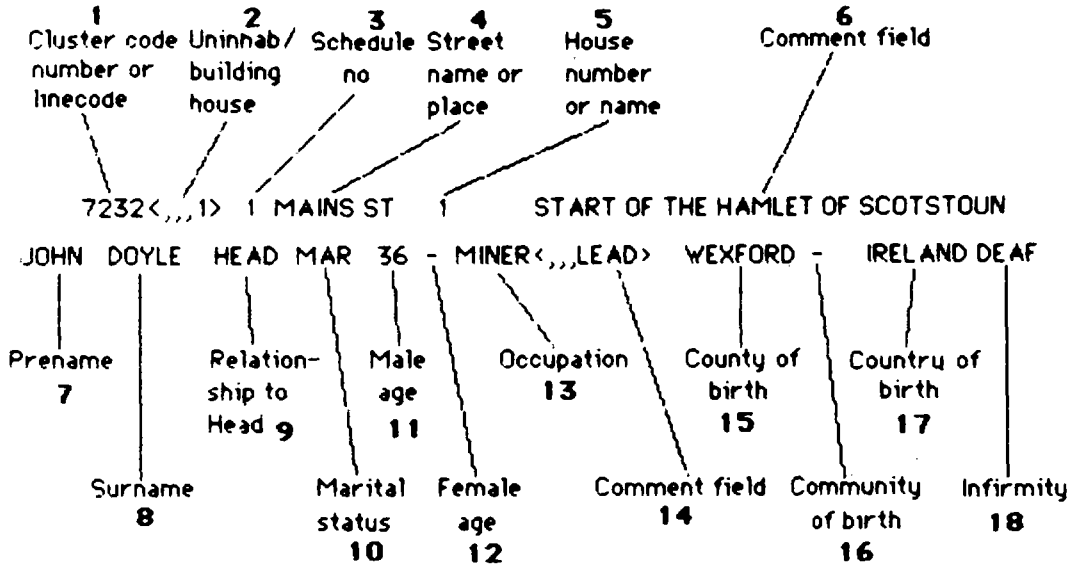
In addition to the data files containing transcripts of the household enumeration, each enumeration book contains certain other information of which the most important is the description, on the very first page of each book, of the area to be enumerated. To facilitate computer processing the basic transcript files did not include this extra information, but at a later date separate files were prepared (for all clusters other than those relating to institutions) which included both the description of the enumeration district and certain basic data which allows the user to relate the individual data clusters both to the original record office reference numbers and to the published census tables.

These files, copies of which are held by the Data Archive, were prepared on the same principles as the main data files; in particular they represent the most faithful reading which the person who did the transcription believed that she could attain without local knowledge but by attempting at all times to record what she believed was written. Details of the content and structure of these files are given in a later section.

5. LAYOUT OF THE TRANSCRIPT DATA

As part of the original checking procedures, the data were processed by a specially written program which tabulated the data relating to individuals into a form roughly similar to the original format of the enumerators' books. The data on houses and households were listed on a separate line preceding each house or household. Data in this format are now available for student use from the University of Edinburgh Economic and Social History Department and also appear in the Chadwyck-Healey microfiche edition of the dataset. Figure 3 gives an example of the data in this form.

The format of the data in this format is shown (schematically and with each field numbered) below, the numbers providing cross-references to the text which follows, which also provides a general introduction to the different data fields. In the tabulated format, each household or person record starts on a new line. Where any entry exceeds the tabulated width available for its field it is allowed to run across into the following fields and the remaining entries for that record are shifted to the line below.



A. Household Records.

1. Cluster code number or linecode: For the first household of each file this field contains the cluster code number. Thereafter, for each household, it contains a number between 1 and 100 indicating one or other of the many slightly different combinations of ways in which enumerators used schedule numbers, addresses, and different length lines to indicate new houses or new households within houses. The precise details of all the different values of these codes will not be of significance for most users; in brief, codes 2 to 49 indicate a belief on our part that the enumerator thought that he had identified a new house according to the complex and contradictory guidelines given to him by the census authorities, codes 51 to 89 indicate that we believe that he was or should have been recording a new household within the same house. A few of the codes have special significance. Thus, a code of 1 indicates the start of a second or subsequent enumerator's book within a cluster, a code between 91 and 95 indicates a new household within an institution, and codes 97 or 98 indicate the creation of a new 'dummy' household for one or other of the groups of people who were identified by the enumerator as living in barns, sheds, boats, tents, etc. (see Section 1 above). Code 100, which occurs rarely, indicates a continuation of the previous household and should be ignored for most purposes; it was used to create a new dummy household for computer processing of very large households which would otherwise have overrun table arrays used by the software.

2. Uninhabited or building house(s). This bracketed comment field occurs wherever the enumerator recorded the presence of uninhabited houses or houses under construction. It is also used for other, similar, information (e.g. 'Family away').

The entries in this field are a direct transcript of what the enumerator wrote, and they appear in the data set on the household record of <u>both</u> the house which precedes and that which follows the empty house or building site, as long as these houses appear to be physically proximate to the site. Commas within the field indicate the location of the site relative to the household, and also whether the entries relate to houses empty or under construction or both. The entry in the example above, which follows three commas, indicates that the next house in the street was uninhabited. A similar entry followed by a single comma would indicate that the preceding house was uninhabited. Entries relating to houses under construction are preceded by no commas when they relate to the preceding house site and two commas if the site is the next down the street.

3. Schedule number. This field contains a transcription of the enumerator's entry in the column in his book headed 'No. of Householder's Schedule'.

4. Street or place name. The enumerator was instructed to record 'the name of the Street, Square, etc. where the house is situate .... or, if the house be situate in the country, any distinctive Name by which it may be known'. He was also required to record housenames or numbers within streets or places. This field contains a transcript of the street or place parts of the address information recorded.

5. House number or house name. This field contains the remainder of the address information not recorded in field 4.

6. Comment. In this field are recorded any comments made by the enumerator about the situation of a particular house or household, together with any notes made by the research staff on problems of transcription. The first few households and the last household may also contain notes by the research staff on the quality of the book or on any other point which they believed might help in interpreting particular parts of the data contained in the cluster.

B. Person records.

7. Prename. A transcription of the first part of the enumerator's response in the column 'Name and Surname' on the schedule. Titles are also recorded in this field except where they are embedded in the Surname,

e.g.: 'EARL OF/LUCAN' or 'LORD WILLIAM/SMITH',
but 'JOHN/SMITH VISCOUNT HUME'.

8. Surname. The remainder of the 'Name and Surname' record not included in field 7.

9. Relationship to head. A transcription of the material which the enumerator had recorded from the schedule column headed 'Relation to Head of Family. State whether Wife, Son, Daughter or other Relative, Visitor or Servant'.

10. Marital status. Transcription of the 'Condition' field from the enumerator's book. Note that enumerators were instructed not to record this information for young children, with the result that this field is usually blank for persons under marriageable age.

11. Male age. In the enumerators' books the ages of males and females were recorded in separate columns and this practice has been carried over into the transcript. As noted in Section 3, all ages were recorded in numerical form (with any non-numerical entries entered in the comment fields). Ages containing months, weeks or days are thus translated into decimal format. Ages of 999 indicate 'Not known', ages of 998 indicate 'Infant'.

12. Female age. The female equivalent of field 11.

13. Occupation. A transcription of the complete entries from the enumerator's book column headed ' Rank, Profession, or Occupation'.

14. Comment. Though shown in the example appended to the Occupation field, comment fields can relate to any person field and most household fields. The procedures whereby Registrars and others checked the books and Census Office checkers marked comments on them was described in Section 1 of this Guide; the extent to which these various amendments and comments were included in the main data transcript has also been discussed above. In all cases where these comments and amendments could be interpreted they were included in the transcript, normally in the comment fields. The number of commas preceding the entry indicate the origin of the comment entry. Three commas, as in the example above, indicate a checker's note. Two commas would indicate an insertion by a Registrar or other commentator at the verification stage, while one comma indicates that the field contains the enumerator's original entry and that the main field contains some amendment ('<, ->' indicates an insertion by the research staff where the enumerator's entry was blank). No commas indicate an insertion or amendment made during the preparation of the data for punching; the commonest cases occur where order has been changed in the birthplace columns or where dittoes occur in visually correct but illogical positions in the original data (see Section 4 above).

15. County of birth. Birthplace information was required to be provided to the enumerator in the following form for a person resident in England and Wales: 'Opposite the names of those born in England, write the County and Town or Parish. If born in Scotland, Ireland, the British Colonies, the East Indies, or in Foreign Parts, state the Country; in the last case, if a British Subject, add "British Subject".' The instructions for Scotland were identical except that 'England' and 'Scotland' were transposed. These instructions were not always precisely followed, especially by persons born in Ireland who, in spite of the instructions, often gave a county of birth. In the National Sample data set, therefore,

the 'County of birth' field contains any information given on county of birth within the British Isles. London (together with entries such as 'Middlesex London') is treated as a county and a few places which were not strictly counties (e.g. 'Isle of Wight' and 'Ulster') are also recorded in this field. Counties or regions of other countries are, however, confined to the 'Country of birth' field.

16. Community of birth. This field contains a transcript of the 'Town or parish' part of the birthplace entries. Sometimes places smaller than parishes are listed, including street names; these are also entered here.

17. Country of birth. Information on country of birth (where specified) is recorded here, together with any more detailed birthplace information given for persons born overseas. The returns required on 'British Subjects' and also such entries as 'At sea' are also included in this field.

18. Infirmity. The material recorded in this field is taken from the 'Whether Blind or Deaf-and-Dumb' column of the enumerators' books. This information is normally believed by modern scholars to be almost useless since the numbers of blind people recorded are very low and information on the deaf is recorded inconsistently. In addition, because of the position of this field in the books, the photocopying process will almost certainly have rendered some entries invisible to the punchers; users are therefore warned, on two grounds, of the dangers of using this material for quantitative research.

File documentation sheets

As noted in an earlier section, these sheets contain basic reference information for each non-institutional cluster. The information is largely taken from the first page of the relevant enumerators' books, supplemented by the record office reference material. Figure 4 shows an example.

Each sheet is headed with the cluster number. Then, for each separate enumerator's book within the cluster a series of pieces of information is repeated, as follows:

'ED NUMBER': the sequence number of the enumeration district within the cluster;

'Classification': the full reference number of the enumeration district as recorded by the Registrar Generals' offices, together with the Record Offices' reference numbers. For England and Wales the Registrar General's reference [RG(E)] comprises the Registration District followed by the Registration SubDistrict, and the Parish sequence number; the Enumeration District sequence number or numbers within the parish are included where they are integral to the heading of the first pages of the original book; sometimes enumeration districts are sequenced by letters instead of numerically. On the next line, the Public Record Office Class [HO 107] is followed by the PRO box number and then the folio numbers of the book. For a Scotland enumeration district there is only a single reference number, consisting of the New Register House volume number followed, where given, by the sequence number of the book within the volume.

'County', 'City or burgh', and 'Parish' are in most cases self explanatory. The 'ED Number' entry is the sequence number or letter of the particular book within the sequence of the parish or place; these do not always appear on the book itself but have sometimes been collected separately by the research team.

'ED Description' is a straight transcript, as far as it is legible from the photocopy, of the description of the boundaries of the district provided for the enumerator.

6. CODING AND STANDARDISATION PROCEDURES

Why was standardisation necessary?

1. The data were punched exactly as they were recorded in the original source (except that upper case only was used, being all that was available on standard card punches at the time that the project began). However, because the data were entered as far as possible unchanged, the 'same' entity can be identified in the data in many different forms, either through abbreviations (e.g., for Leicester- shire, as 'Leic', 'Leics', Leicestersh', etc) or through misspellings (e.g. 'Liecs', 'Lestershire', etc). All the different data fields are subject to both these problems to some degree. Thus, even if all that is required is a straightforward count of the numbers of people possessing any single characteristic it is necessary to find some way of consolidating all variants of that characteristic into a single data category. Standardisation is a convenient means of doing this.

2. In the case of the occupation field, a single entry contains multiple information fields, but not in a standardised format. Thus, for example, one entry may read 'Farmer of 45 acres employing 2 labourers', while the next may read 'Farmer 45a 1 lab indoor and 1 outdoor'. Further on, an entry may read 'Farmer emp 45a' followed later by 'Coal merchant and Farmer of 45a'. There is a wealth of information in such data fields and it is necessary to find some standard way of holding the various components (first and second occupations, acres farmed, number of employees, etc) in a data set used in analysis.

3. Natural language transcription gives data fields of very variable lengths (e.g. 'Lab' and 'Servant lately in the employment of Sir T.F. Shelley she is not in my service but staying in the house'). Such data fields are difficult if not impossible to process by most standard data analysis packages. More efficient processing can be achieved by having the data transformed into a fixed format and preferably a numerical form.

4. Except at an initial stage of operations, it is frequently very inconvenient to have data fields containing values occupying a very large number of categories since, even in single variable analysis, large numbers of categories are difficult to assimilate and in multi-variable contingency analysis the results become almost impossible to read. The sampling error on each of a very large number of ungrouped categories can easily be so large that interesting relationships are missed and false inferences are made.

This means that there are distinct advantages in being able to group data values in ways which allow responses which are 'close' to each other to be handled together (e.g. where appropriate to be able to group 'servant' and 'cook' and 'lady's maid" and 'butler' together to be handled differently from 'shepherd' and 'ploughboy' and 'cattle man'). If data values have been transformed into some standard numerical form then this makes it possible to locate characteristics which are 'similar' to each other close together on the numerical scale, and characteristics which are 'dissimilar' further away.

5. As soon as one begins to contemplate this process, however, a fifth reason for standardisation becomes apparent. Data values are not analytically meaningful in themselves but only in terms of the underlying concepts of which they may be taken as indicators. This is particularly clear with occupational titles, the analytical significance of which depend crucially on the underlying phenomenon which is of interest at any particular point in time. Thus a single occupational title can tell us something about its owner's employment status, skill level, training requirements, nature of workplace, level of remuneration, status in the wider community and many other things. When one comes to consolidate occupational titles into larger groups it is thus crucial that one does this in different ways according to the underlying phenomenon of interest. This in turn means that occupational data values have to be mapped onto multiple continua, with different classification systems relating to different underlying concepts. A similar issue can also arise with data on community of birth (which reveals distance of migration, size of community of birth, occupational structure of this community and so on - and might then imply that a single birthplace title could be classified in a number of different ways depending on the topic of most interest at a particular point in time).

Coding directories

In this project all standardisation of the data was done by the computer, using specially created coding directories (or dictionaries). As each data file was processed, all the relationship to head of household strings were checked by the computer against a relationship directory and a file containing a list of currently uncoded relationship strings was printed. A code was then edited into this file against each of the previously uncoded strings and the whole directory was then updated so that it included the new strings. The data file was then re-processed by the computer which tagged each relationship string with the appropriate code. The same procedure was then undertaken for the next data file until all files had been processed. Similar directories were created for all

counties and countries of birth, and for all marital status strings. A proportion of all occupational strings were coded in the same way. The resulting files, which currently contain codes for about 150 different marital status strings, about 2200 different strings referring to relationship to household head, almost 2900 different county of birth strings, and over 1000 country strings (as well as over 13000 occupational titles), represent a major investment of time and effort and are potentially of use in many other census research projects. They have therefore been deposited in the ESRC Data Archive at the University of Essex. Brief documentation is also available from the Archive on the different files and on the principles used to determine the codes allocated to any data response.

Standardisation procedures

As indicated above, standardisation may be adopted for many reasons of methodological significance as well as of convenience. However, it is important that flexibility and sensitivity of analysis is not thereby unnecessarily lost. For these reasons, in this project the codes and procedures adopted have been quite complex. In particular:

i) Every discrete characteristic appearing in the data has been given its own unique numerical code; this means that subsequent modification of derived codes and possible groupings and recombinations can be effected at any time without returning to the original character strings.

ii) Because the programs accessed a common dictionary for all clusters it was not normally possible at the coding stage to take into account regional variations in the use of terms. This is mainly a problem with occupations. However, in the final version of the Public Data Format 2 files (see below) some modifications to the standard codes have been made manually where the results would otherwise have been seriously misleading.

iii) Numerical codes have been mapped onto standardisation continua in such a way that 'similar' character strings from the original data are grouped relatively near to each other and 'identical characteristics are given the same code (thus, for example, while 'brother' is coded 3011 and 'sister' is coded 3211 and 'cousin' is coded 3811, 'grandson' is coded 4011, 'head's grandson' is 4014, 'son's son' is 4015, 'daughter's son' is 4016 and 'son-in-law's son' is 4072; 'granddaughter' and 'grdaughter' and 'gdaur' are all coded 4111, while 'daughter's daughter' is coded 4116' and 'adopted granddaughter' is 4153).

iv) Standardisation always involves inference about what was really intended. Sometimes this problem is made worse by illegibility, perverse spelling, or failures of communication between census respondent and enumerator or between enumerator and coder. Possible difficulties in inference have been flagged in the coded data by the inclusion of 'inference codes' for all the major variables. Details of the codes are given in the published documentation of the respective variables.

v) Coding of the many thousand differing responses was a major task and there were occasions when judgement had to be used to decide on the appropriate category into which a particular item of data should be placed. Coders used their best judgement, aided by a range of reference books. It is clear however, that occasionally the codes allocated will be wrong, more often there will be some doubt about what is appropriate. This is particularly true for occupations which, because of their complexity, were coded onto multiple continua, full details of which may be found in the documentation. Users of the data may from time to time wish to reallocate particular responses; this they should normally be able to do without great difficulty since in the Public Data Format versions of the data (see below) raw transcript as well as coded data are always present in the file.

Coding of community of birth

In the initial stages of the project, coding of community of birth was performed by procedures similar to those used for other data fields; in particular, a directory was prepared and coding was done on multiple criteria, including information on grid reference, population and the occupational structure in 1831. This proved very time consuming and also impractical because of the large number of places which shared a name with some other settlement.

Later data sets have been coded on grid reference only, using information from the machine-readable Ordnance Survey Quarter Inch Gazetteer which we mounted in Edinburgh and accessed interactively by a specially written program (GAZ). This program and its associated database are now available to all users of the Edinburgh Computing Service since it has been taken over and is now maintained as a general user facility as part of the Scottish Data Centre. In a proportion of cases where entries were not found in the Quarter Inch Gazetteer they were searched for manually in other gazetteers or on the OS One Inch maps but by no means all could be found in the time available. Particular problems arose with parishes within cities where many entries are completely ambiguous. Here, as elsewhere where major ambiguities were present, no attempt has been made to 'guess' the most appropriate response.

7. PUBLIC DATA FORMAT FILES

The Public Data Format File concept

From the outset, the National Sample data was intended to become a public data resource. It was thus essential that we should consider how best to develop a data format which would most readily facilitate the portability of the data across different machines and between different users. Experience on other projects eventually led to the development of the 'Public Data Format' which has subsequently been implemented in two forms, a 'Public Data Format 1' (PDF1) in which all the data has been cast, and an enhanced version (PDF2) in which at present only the National Subsample Files are available.

Full documentation of the PDF1 and PDF2 files is available from the Data Archive but the latter, in particular, is a bulky document so it may be useful to potential users to have brief descriptions of each of the formats here.

Public Data Format 1 files

The basic characteristics of a PDF1 file are shown in the accompanying example (Figure 5) and in the diagram below; they may be summarised as follows:

All data is stored in upper case and only a limited number of special symbol characters is permitted; this ensures maximum portability.

Each line has a maximum length of 80 characters; 'card-image' files of this kind are by far the most readily portable.

The data are layed out in the file in a standard fixed field format; this occupies more space than would be the case with a free format representation, but has the advantage of general portalibilty without any possible software dependence, and provides a data set from which it is easy even for relatively naive programmers to extract material as required.

Transcript and coded data are both included in the file, though the bracketed information fields (which in practice only rarely contain material of interest) have been excluded in order to lessen the space demands of the final product.

Each line has a numeric record identification code in columns 1 and 2.

Header information is found on linetype 11
Household information is found on line types 21 to 29
Locational information for each household is on line types 31 to 33
Transcript information for each individual is on line types 41 to 49
Coded versions of transcript information is on line types 51 and 52

FORMAT OF A PDF1 FILE

LINE **CONTENTS**

NB. The numbers in parentheses indicate the number of columns which the field occupies in the record.

11 |Number of households in file (4), number of individuals in file (5)

21 |Cluster code (5), No. of individuals in household (3), New household indicator(4), Schedule no (66)
22 |Information on whether neighbouring houses were unoccupied or building (78)
23 |Place of residence (78)
24 |Address of residence (78)
25 |Other information about household (78)
26 | Ditto (contd)
27 | " "
28 | " "
29 | " "

31 |Location of cluster (78)
32 |County within which cluster is located (78)
33 |Grid reference of cluster (8), grid ref inference code (1), county code (3), place type code (2)

41 |Sequence number of person in household (2), Prename (76)
42 |Surname (78)
43 |Relationship to household head (53), Marital status (25)
44 |Age if male (7), Age if female (7), County of birth (64)
45 |Occupational title string (78)
46 | Ditto (contd) (78)
47 | " "
48 |Country of birth(78)
49 |Community of birth (40), Whether Blind, deaf or dumb (38)

51 |Codes for relationship (4+1), marital status (2+1), county (3+1), country (4+1), community (8+1)
52 |Code for occupations (59)

Details on each additional member of the household would then follow on a further set of lines 41 to 52

The next household would then follow on lines 21 to 33, followed by further individuals and then further households

A '9' in column 1 marks the end of the file.

One other important feature of the PDF1 files should also be noted. While the Data Archive holds the complete machine-readable coding directories, it should not normally be necessary to use these directories; the codes used in any individual PDF1 file are self documenting since for each entry both the character string and the coded version are present.

Files in this format were mainly created in the years between 1981 and 1984. The entire National Subsample is available from the Data Archive in the full version of the PDF1 format. Every cluster of the whole main sample is also available from the Archive, but in a restricted version of the PDF1 format in which there are no codes for occupations and no grid references for places of birth. The main sample files use the same numbering system as the transcript files but with the prefix 'F' (e.g. F1011). Where very large files were involved they were subdivided for processing and these subdivided files are then indicated by suffixes 'A' and 'B' (e.g. F0608A and F0608B).

Except for institutions, the National Subsample PDF1 files are numbered in the same way as the basic subsample files described above, but with the prefix 'FX' (e.g.FXA70). However, for convenience of handling, in some cases institutions from different Census Divisions are grouped together: all London institutions are in files numbered FXA09-FXF09, and all Scottish institutions are in files numbered FXA74-FXF74; all workhouse samples from Divisions 2, 3, 4 and 5 are in files numbered FXA93-FXF93; other insitution samples from these Divisions are in FXA94-FXF94; workhouse samples from Divisions 6 through 11 are in FXA98-FXF98; other institution samples from these Divisions are in FXA99-FXF99.

Public Data Format 2 files

The PDF1 files provide an excellent data set when the only topic of interest is individuals and their households, but their essentially simple logical structure limit exploitation of the full richness of the original census data. As a result, between 1984 and 1987 a further set of software was developed to transform PDF1 files into a much enhanced format, PDF2. The most important new feature of the PDF2 files is that an extra 'level', the family, has been inserted in the data. As a result, not only do individuals now belong to households but they are also members of one or more families. Moreover, families are formed not only when they are directly related to the household head but also wherever family relationships may be found within the original data (thus lodgers' families, families of servants, relatives and even visitors, are all created wherever possible). In addition the dataset has been enriched in a number of other ways. Indicators of household composition, family composition, family life

cycle stage, family migration status and many other structural attributes are now included within the file. Counts are made of numbers of people in households and families and of the numbers possessing a whole range of attributes (age, relationships, marital status, occupations) and these counts, as well as some information on significant individuals (heads, wives, eldest children, for example) are distributed to a number of different levels within the file (e.g. each individual has a great deal of information about other members of the family and household attached to his or her own person record). A number of key familial relationships between individuals are also identified and stored within the file. A number of summary indicators are included (for example occupational titles, in addition to having an occupationally specific coded value present, are also summarised into occupational orders and into socio-economic groups).

The basic structure of a PDF2 file is closely related to that of a PDF1 file (and is equally self-documenting) except that the substantially increased number of field types means that a three column line identification code has to be used. Instead of there being just three levels in the file (cluster, household and person) there are now six (cluster, enumeration district, house, household, family and individual) - though the complete software to produce enumeration district level data has not been implemented and house level data is not normally produced in the National Subsample files, which are the only ones at present in full PDF2 format.

PDF2 files will be available from the Data Archive for the entire National Subsample from early 1988. Some additional main sample cluster files may also be produced from time to time and the Archive will have details of any that have been deposited with it. All PDF2 files deposited in the Archive will have been subjected to a number of manual and computer-based validation checks. As a result, the codes and some features of the raw transcripts will not exactly correspond with earlier versions of a particular file. The National Subsample PDF2 files are grouped by stratum, division and replication in the same way as the PDF1 files. They are prefixed by 'PX'. For some purposes the code numbers are converted to an all-digit format. Thus, for example, FXA70 becomes in PDF2 PX7011, FXB70 becomes PX7021, FXC71 becomes PX7131, and FXD72 becomes PX7241.

Postprocessors

In order to facilitate the use of PDF2 files in modern relational and other database management systems, postprocessors exist which will convert PDF2 files into a series of flat files which include appropriate

relational keys; they will also produce files which may be entered directly into a general purpose SIR schema. Potential users should note that due to inadequate staff resources for full checking the resulting SIR files are not guaranteed to be free from error in use.

8. SOFTWARE

Most of the software used during the early phases of this project is of little use to users outside Edinburgh; some is already obsolete. In the early 1970's when the project started there was little in the way of package database software available, and an extreme premium had to be placed on processing efficiency when dealing with datasets of the size that we were confronting. This meant that a great deal of purpose-built software was developed, much of it exploiting particularly advantageous features of the software environments available to us at that time.

As noted above, the first data structuring software that was used was GENDATA, developed by the Cambridge Group, written in PL360 and designed to interface with the Newcastle File Handling System. We later used an OS version of this software written in Cambridge, and later still wrote our own emulator of it using the language IMP (LINDATA). IMP was also used for all the directory handling and coding software and for the various versions of the tabulation software. Some of this software is written in older generations of IMP and some work will probably be needed to make it available even in Edinburgh after July 1987.

The CENDEP software which converts PDF1 to PDF2 files, is, however, written in highly portable FORTRAN and a copy of all the programs has been deposited in the Data Archive.

9. SUMMARY OF AVAILABLE DATA

1. The following machine-readable material is available from the ESRC Data Archive, University of Essex, subject to their usual terms and conditions:

a) Raw transcripts of the clusters from the original main sample.

b) PDF1 files for all main sample clusters; these files are not coded for occupational titles and community of birth.

c) PDF1 files for the National Subsample, completely coded.

d) PDF2 files for the National Subsample (from February 1988).

e) Coding directories for the main data fields, updated to late 1984.

f) CENDEP software and associated postprocessors.

g) A list of all sample locations, including grid references.

2. The Data Archive also holds documentation on the datasets which it distributes. In addition, the Archive holds a hard copy of the File Documentation Data which gives the source references and boundaries of each data cluster. In due course these documentation sheets may also be available from the Archive in machine-readable form.

3. Single clusters or groups of clusters in QUEST-compatible formats for use with BBC micros are available for educational purposes from the Edinburgh University Data Library Service.

4. The complete data set in a tabulated format approximating to the original layout of the data is published in microfiche by Chadwyck-Healey.

5. Printouts of individual data clusters in the same format may be purchased from the Department of Economic and Social History at the University of Edinburgh.

10. DOCUMENTATION

Further detailed documentation is available as follows:

a) An overview of the development of the project and detailed accounts of the processes involved in generating the data set in its different forms can be obtained by reading the Final Reports on the periods of SSRC/ESRC funding. These are available from the British Lending Library. The first project was entitled 'Preparation and analysis of a machine-readable national sample from the 1851 census of Great Britain' (HR 2066). The second project had the title 'Establishment of the 1851 census national sample as a data library' (H00230016). The third project was called '1851 census national sample data library: software implementation phase' (H00232032).

b) Full documentation of the PDF1 files and of most of the key variables is included in the Final Report on the second project. A fuller account of the principles and practices of occupational coding has also been prepared and it is intended that it will be published in an appropriate journal and a copy be deposited in the Data Archive.

c) Complete documentation on the PDF2 files and on the associated CENDEP software is available from the Data Archive.

d) The accompanying material for the Chadwyck-Healey microfiche edition of the data includes a list of the clusters, the complete set of file documentation sheets, and a summary of the dataset and some possible uses which may be made of it. The microfiche is an excellent reference document for anyone needing to examine the context of individual households from the National Subsample Files.