

PLANNING FOR THE 2001 CENSUS OF THE UNITED KINGDOM

Alex Clark, Deputy Director 2001 Census

1. Introduction

The next Census of Population and Housing in the United Kingdom is scheduled for 2001. This paper sets out the major milestones on the path towards 2001, describes the research which is being carried out to identify changes in demand, reviews business process re-engineering of the methodology used to collect, process and disseminate data and also gives an insight into the potential for contracting out aspects of the work to the private sector.

It includes a description of the consultative machinery that has been set up to enable a two way dialogue with customers to ensure that the main users of census data contribute effectively to the content and design of the census. In particular, it will outline the opportunities for using digital mapping, boundaries and a geo-referenced address gazetteer as the base for census geography, both to design data collection areas, produce maps and address lists for enumerators and to manage the operation. It goes on to review work being done on changing collection methodology to reduce costs and help deal with enumeration problems, to highlight developments in optical scanning, to identify the opportunities for automatically coding responses, a new way of imputing missing data and first thoughts on remote access to data. It will include a summary of a recent review into the opportunities for contracting out more of the work connected with collecting and processing data to commercial undertakings, indicate how this work is being taken forward and how the associated risks are being managed.

2. Building a model which will meet user requirements.

2001 Census methodology will build on that used for 1991, incorporating alterations to meet changing user requirements, changes in society and improvements since the last Census. Feedback on 1991 Census performance comes from a variety of sources. Regular contact with users over queries, problems and provision of support services as output is produced and used, is a primary source of information about individual products. Internal reviews of the census projects for 1991 provided information about performance from project managers. The formal statistical checks incorporated in the Census Validation Survey¹ and other internal demographic checks have provided coverage and quality assessments of the data. Independent studies such as the 'Estimating with Confidence programme'² will add to this. The recently completed 1991 Census user debriefing survey sought information from users throughout the country, via a questionnaire and a series of regional meetings. Most of this is covered in the official review of the 1991 Census³.

Census advisory groups, which cover the main users of census information, are helping to co-ordinate and define user requirements. Special reviews, such as the 1994 Association for Geographic Information/Census Offices roundtable⁴ and the Royal Statistical Society/British Society for Population Studies⁵ conference on looking towards the 2001 Census, demonstrate a healthy interest in providing user needs. User working groups are provided detailed comments and helping with the research into user needs, data collection and the definition of outputs.

3. 2001 Census Aims

The four high-level strategic aims for the 2001 Census are:

- to ensure that the question content is appropriate to meet the demonstrated requirements of customers, taking account of consideration of value for money;
- to deliver products and services to meet legal obligations and customer's needs within stated quality standards and to a pre-defined timetable;
- to ensure that all aspects of the census collection operation, and the dissemination of results, are acceptable to the public and comply with Data Protection law;
- to demonstrate that the census represents value for money.

A research and development programme running from 1992 to 1999⁶ will seek improved performance and increased value for money for both the Census Offices and users. This will be done by considering ways of reducing development costs by increasing efficiency and harnessing new technology. Efforts will be made to reduce operational costs by considering how changes to methodology might result in a reduction in the number of temporary staff used to collect and process the data. A key aim will be to produce better value products by meeting quality, timeliness and coverage targets. Ways of improving the marketing of information, widening the customer base and considering joint business ventures will all be addressed as part of a more focused marketing programme.

One of the planning assumptions for the census is that the aim should be to have no real increase in costs over the 1991 Census. This can only be done by a combination of efficiency savings and increased cost recovery or by reducing the scope. Additional income could come from a better range of products, a wider customer base, but also from a review of the charging base and possible increases in prices.

4. Key milestones

The key milestones in the plan are:

1993	Statistical Policy - set standards (done)
end of 1993	Agree testing strategy and programme (done)
1994-96	Small scale testing programme (started)
May 1995	Define small scale testing programme and start cognitive research with form fillers (done)
June 1995	Agree key IT milestones and dates for defining IT strategy for 2001 (done)
Sept 1995	First question wording tests (complete)
Dec 1995	Identify 1997 Census Test areas (done)
June 1996	Agree question set for 1997 Census Test
1996	Set Legal Basis - including European Community Requirements
1996	Review strategy for provision of statistical information on population and housing post-2001
Jan 1997	Start user consultation on business cases for 2001 question set
April 1997	1997 Census Test major operational field trial
late 1997	Prepare outline specifications for Data collection and data processing Prepare outline strategy for data dissemination
mid 1998	Agree topics for census questions Settle marketing strategy Publish Plans for 2001 (Government White Paper)
1999	1999 Census Test: Dress Rehearsal
end 1999	Finalise Data Collection Procedures Census Order laid before Parliament■
early 2000	Dress rehearsal (Processing) Make Census Regulations■ Finalise census validation
2001	Next Census
2001-2003	Produce results to defined timetables
May 2002	Produce 100% basic data counts for Government Resource allocation work.

■ Legal base

5. Outline of major activities

a. Data Collection and geography

Although the final selection of topics for inclusion in the 2001 Census will not be made until 1998, it is important that questions are not discounted at that stage because of inadequate early testing. The small scale testing programme which is an incremental step on the path to the first major field trial in 1997 is designed to identify changes in user requirements for topics and then to test any selected new ones or changes to existing classifications in order to define a set of questions for 1997 and then for the final selection in 1998.

The programme is using cognitive research⁷ to discuss concepts and possible questions with a panel of users and then is feeding the results in a question testing programme, using trained interviewers to carry out assessment of questions. A series of small tests has been completed and the project team are now working towards devising a set of questions for the 1997 Census Test. A small scale evaluation follow-up will take place after the Test to determine whether the questions worked in practice.

It is likely that an income question will be included in the 1997 Census Test, but it has not yet been decided whether it will be addressed to individuals or be asked of the household as a whole. Other developments, include a possible change to both household and dwelling definition, a question on carers and changes to existing questions to widen classifications, and the introduction of a number of minor new questions. So far, it is providing difficult to get users to agree to drop existing topics to make way for new ones and, as one of the objectives for 2001 is not to increase the load on respondents, some tough decisions will need to be taken.

The geographic base for 2001 is likely to be based on individual addresses or postcodes. This is a change from the enumeration districts used up until 1991 and is being introduced to meet users needs, make efficiency gains and to make use of the availability of a number of digital products which have been introduced by the Ordnance Survey. The system will use digital boundaries for both 1991 Census enumeration districts and current statutory areas, electronic maps in both raster and vector form, the Ordnance Survey address gazetteer which contains a 1 metre reference for properties and ARCINFO software. A number of pilot trials have demonstrated how these inputs can be used to create enumeration areas for 2001. For the 1997 Census Test the collection areas will be based on 1991 enumeration districts with changes for housing growth and boundary changes to statutory areas. For the first time in England and Wales, collection geography will be divorced from output geography. The introduction of the individual address references enables output areas and digital boundaries to be created automatically after processing. This will give users a more flexible geography which can be defined at the time of producing output and not pre-determined some years earlier.

The geography system will use the address gazetteer to identify new housing and addresses affected by boundary changes and, in this way, will enable planning software to create new enumeration districts interactively - a full automatic system is a possible later development. The system will also print address listings for enumerators and will create enumeration districts on a single sheet of paper

at one scale. These are exciting developments which will not only save a considerable amount of money, but will provide a better service to data collection staff and eventually to users.

A major change in data collection methodology is being tested in 1997. Forms will still be delivered by enumerators, but in half of the areas being tested, completed forms will be posted back to local Census Officers who will control follow-up. In the post back areas, the size of enumeration districts have been increased and first estimates are that the number of enumerators for a full census could be reduced from 100,000+ to 60,000+ and then to a much smaller number for follow-up. One of the key objectives of the postal system is to release resources to enable more effort to be put into dealing with under enumeration. A number of options to achieve this are being pursued, but as the 1997 Census Test will be a voluntary one, it will not be easy to test them thoroughly.

Two different types of questionnaire will be tested in 1997. A conventional matrix form will be compared with a page per person approach. Some work is also being done into investigating whether there is a significant correlation between under coverage in small areas, as identified in the estimating with confidence programme, and the intelligence available in the data collection management information system. A link would give the opportunity to target the worst areas during the census, rather than report results afterwards.

b. Processing system design changes

For 2001 a number of basic changes are being considered to enable a much more flexible system, based on individual forms, to be considered. Firstly, the method of identifying dwellings might be simplified to enable the data to be captured by a question on the form with forms for a dwelling being linked together by an address gazetteer built into the processing database; second, a free flow system of processing individual forms is being tested because this will be able to cope more easily with the proposed postal return of questionnaires; third, scanning is being investigated to produce images and capture some data automatically; fourth, coding will be automated for answers to the basic demographic questions and work is being done to assess the viability of automatically coding the 'hard to code' questions, like occupation; fifth, neural networks are being considered as an alternative to the existing automatic edit imputation processes.

These changes are described in the detailed sections of the paper.

Data Capture

The aim is to develop a cost effective and fully integrated data entry strategy to optimise and streamline capture, coding and editing of data.

A number of different ways of tackling the issue are being pursued. Fully automatic capture, using scanning, OMR and OCR technology to capture data with clerical intervention to capture 'problem' data has been rejected because it is highly unlikely that a reliable system will be available to be used for 2001.

Various ways of enhancing the traditional approach are being considered, including using coding staff to code and capture the data in one operation. However, the most interesting option introduces the use of image scanning in a combination approach, where forms are scanned to capture part of the data using recognition software and to hold an image of un-captured data which would be keyed/coded interactively or keyed and coded automatically.

So far, investigations into the potential of scanning census forms to capture basic data are showing promising signs. A mixed use of Optical Mark Reading and Optical Character Recognition techniques presents the possibility of scanning images in combination with using OMR to capture tick box answers and to identify write in answers for further processing, with OCR being used to capture numerics and possibly postcode information.

There are a number of basic design issues which will be addressed in parallel with technical trials:

- Automated Workflow software and organisational aspects of the data capture strategy to optimise the flow of work in an automated environment.
- The scope to increase the number of closed questions to minimise write in answers and to capture a mix of tick boxes and written responses.

Coding of data

Issues that will shape the 2001 coding design are:

- The impact of the data capture method on processing. The approach to coding will vary depending on the data capture method and the stage at which data is captured.
- The interface between coding and other processing activities. One of the main aims for 2001 is to develop a processing strategy which integrates data capture, coding and editing in a streamlined operation. Rationalisation of activities together with the combination of query resolution and interactive editing are key components.
- The viability of restricting the number of open-ended questions and the scope for increasing the number of tick boxes on the form. Savings made from automating some or all of the 100% basic questions may provide the opportunity to increase sample sizes for the 'hard to code' items. Opportunities to collapse classifications may also exist, depending on the customer requirement.
- The stage at which validation and edit checks take place. Automatic coding allows earlier checking of data but the most cost-effective approach needs to be explored.

The research done to date has demonstrated the feasibility of incorporating

autocoding into the processing system.

For simple classifications automatic coding outperform manual coding in terms of quality and consistency with acceptable match rates which can be improved by enhancing base indexes and increasing the number of tick box responses on the form.

Two address processing systems have been tested with limited success, but it is important to note that no proprietary software will meet census requirements without customisation. Also, the success of automatic coding of addresses is dependent on careful structuring of the way that address data is collected from respondents.

Complex classifications pose a more difficult problem, but packages and reference files exist for two of the most difficult classifications, industry and occupation. A sample of records taken from the UK Economic Activity sub-sample has been tested using one of these packages, CASOC. Nearly all responses were coded but only 61% of cases matched 1991 codes. Further work is being done, including an assessment of the benefits of incorporating a neural network in one of the packages.

Editing

The choice of the Editing and Imputation systems and processes will depend on evaluation of options which demonstrate the ability to meet requirements and find the right balance between complexity, flexibility, quality of data and cost. The methods chosen need to be practical and statistically sound, will not introduce bias or distortion in the data and meet pre-determined Data Quality levels.

The options being reviewed range from developments of the system used for 1991 and a total rethink to consider the potential of neural networks to provide missing answers.

Imputation Trial using Neural Networks.

A proof of concept study to investigate if neural technology⁹ can successfully impute missing items in census data has been completed.

The results are promising. The next step is to compare this imputation methodology with other more conventional techniques, and with the hot deck approach which was used in the 1991 Census, and to extend the trial further to areas not covered in the proof of concept approach.

Unlike traditional computing approaches which need to be explicitly programmed, neural computing techniques automatically learn solutions from data describing the problem. A neural network can be taught and can learn about personal and household profiles provided in the census data in order to accurately impute missing values.

The neural network will go through this learning process, commonly known as

training. By using analysis tools, a model which has learnt profiles from the data may be analysed to show the relationships it has learnt.

Initial conclusions are that the concept that neural networks can impute census data, without introducing bias or distortion, and accommodating the particular census concerns is feasible. Some further evaluation of the method against the 1991 Hotdeck process and other imputation techniques will be carried out in an extended trial to examine how the technology could be incorporated into a live processing environment.

Data Quality

Most of the criticism of the 1991 Census centred round the quality of data, and although the prototyping approach being used for 2001 will help, it is still necessary to have a data quality system.

A data quality strategy can only be effectively implemented if there are methods in place to both *assure* and to *assess* quality.

The aim of a quality control procedure is to detect and measure errors in the different phases of the process. This should be done in such a way to enable corrections to be made, interaction with other processes assessed, and the overall effect on the final result measured.

Quality plans for 2001 are being built around an informal use of Total Quality Management, building on current best practices, and looking to conform with customer requirements, not just specifications of individual products and services.

Within processing there is more scope for more extensive quality assurance, aside from the quality assurance of detailed specifications of computer systems. Data expert team(s) will be set up in the processing office(s) with clearly defined terms of reference and authority to propose changes to the system and take remedial action. Within these teams, topic experts will be developed, with a specialist knowledge which will enable anomalies to be spotted.

The data expert teams will need tools to aid them. This implies a need for some form of *Data Quality Monitoring system*. The aim in implementing a Data Quality Monitoring system would be to provide statistical and management information tracking tools which would:

- Provide timely representation of census results during processing.
- Enable comparisons to be drawn with expected results.
- Highlight trends and bias in the data.

There is a general need to spot data problems earlier. The Data Quality Analysis reports need to be produced in parallel with production.

Current thoughts for a Data Quality Monitoring System are:

- Instantaneous monitoring and reporting of census processing;

- Flexibility - not all reports can be pre-determined;
- Ability to allow comparison with external sources such as administrative records;
- Integral warning systems to alert to problems/trends;
- Indication of the level of imputation being carried out, by area and variable;
- Pattern recognition facility - alerting the Census Offices to unusual patterns in the data;
- Able to provide information on progress - i.e. number of forms processed that day, number of missing or inconsistent items found showing frequency counts/distributions;
- Screen based, with a report facility;
- User friendly and in plain English; and
- Analysis tools to assess the impact of changes, for example to the edit or imputation routines.

It is unlikely that all these requirements could be met by one system. Some may not be possible at all, but the UK Census Offices are conscious of the need to improve on the limited system available in 1991.

c. Outputs

Output from the Census is designed to meet user needs, but is constrained by the available technology, the law governing the release of information and by conventions which protect the confidentiality of information about individuals. Plans for 2001 include work on how the greater flexibility being demanded by users can be met within the need to increase the quantum of income, government guidelines on pricing and sale of products and the limitations of the Census Act and the Data Protection Legislation.

The type of service to be provided will need to be reviewed with a greater emphasis being placed on providing a flexible user driven service, support, consultancy and training service and an increasing number of access paths.

Options to protect income throughout the decade, to give greater predictability of income by entering into service level agreements with users, by changing the current charging policy to levy a charge on the actual number of users or uses, to protect Crown Copyright and possibly police it more rigorously will all have to be considered at the appropriate point in the planning programme for 2001.

The census serves a very wide customer base and that the needs of the unsophisticated, one-off user will have to be catered for as well as the highly expert user with access to the latest technology.

The simplest option for 2001 is the "no change" option of repeating the 1991 Census programme. The absolute "no change" option is a non-runner because there is already evidence that it will not meet the requirements of users in 2001, and technology will have changed.

There are many other options, and the final system will not be identified until later in the decade. The amalgamation of Office of Population Censuses and Surveys and the Central Statistical Office to Office for National Statistics in April 1996 will also introduce new requirements and interfaces.

The basic assumption underlying the current plans is that the output programme will be developed to meet users' needs, but, at the same time, will be a cost-effective approach for the Census Offices, will produce an increase in the quantum of income and give value for money to the government for the high level of investment in the Census.

Output is being planned round a framework which will result in a real increase in income over 1991. The volume of published material will be reduced and an electronic library will be used to simplify access to statistics. Electronic output will be available in the appropriate multi-media formats which will be compatible with the mass market technology of the day.

Access to Census statistics will be improved in two ways. Firstly the Samples of Anonymised Records introduced in 1991 for the first time will be continued and considerations will be given to extending them to enable larger data sets to be available to users. Second, work will be done on developing a user confidentiality access screen which would allow direct access to statistics or data whilst providing sufficient confidentiality safeguards. Remote tabulation software will enable professional users to design their own tables. Further work will be done on how to simplify the confidentiality measures used to protect statistics for small areas. The Census Office is working with Statistics Netherlands on a new approach to this problem, but will also continue to evolve its own solutions.

The emphasis of the Census Office's sales effort will be expanded to increase servicing, training and providing consultancy support to users.

A number of changes will be made to the business arrangements for purchasing statistics and other products. The possibility of establishing long term agreements for large users to give predictability of income will be pursued. This will be coupled with consideration of whether users have an interest in receiving other statistics in the intercensal years to provide an ongoing picture of demographic development - although none will be at as small a geographic disaggregation as the census.

The conditions of sales for Census Statistics will be reviewed, with an aim to protecting copyright and maximising income from royalties.

The Census Offices will be looking outwards to review options for production of joint products with commercial or other partners and to maximise the business arrangements with the existing third party commercial agencies supplying census

statistics and services to the geo-demographics industry. This will include the creation of mixed product sets and the concept of a one stop shop for government statistics.

6. Outsourcing

Funding for the Census is likely to come under scrutiny and is already included in central efficiency assessments which affect all government departments. Although a considerable amount of work was contracted out for the previous census, and a large temporary field force was recruited, the Census Offices managed the work centrally and through its own temporary local managers. A review into the potential for increasing outsourcing was carried out in 1994 and 1995.¹⁰

The review confirmed that it would be inappropriate to outsource the whole census, or even the whole of the field operations, on a contractual basis, since the risk involved and the lack of opportunity for a "second take", made it essential that the Census Offices retained control. These results were encouraging and are in line with the planned approach.

The scope of the census and the wide range of tasks to be undertaken make a total in-house approach inappropriate, since some of the skills needed are not available in-house in sufficient quantity. Parts of the census were outsourced in 1991 and this policy will continue for 2001.

It has been agreed that major additional census activities should be considered for outsourcing

- a. the supply of computing services,
- b. the coding of census forms and their conversion to electronic data,
- c. and the outsourcing of the supply and payment of enumerators for field operations.

Estimated cost reductions were calculated for each of these activities. The unique one-off nature of the census and the peaks of workload that characterise it, do not make it ideal for generating outsourcing cost reductions. Nevertheless, a gross reduction of around 11% of the value of the items selected for outsourcing might be expected and, expressed at Net Present Value (NPV) a net reduction (after adjusting for the costs of outsourcing) averaging 5% on these services may well be realised.

The Census operation is very complex and vulnerable to threats and there is only limited opportunity to recover from disasters. Also as the census carried out under statute, aspects of it are very date critical.

Outsourcing poses risks which can only be managed if:

- A clearly defined specification of the work can be drawn up.
- The capacity and capability is present, within the customer's organisation, to regularly monitor the quality of service provided and, if necessary initiate

corrective action.

- The size and nature of the work package and the timescale for its completion allows a contingency plan to be operated if necessary, using appropriate alternative skills.

Criteria were developed to reject candidates for outsourcing. These are too detailed to be described fully in this paper but include the need for intimate census knowledge which could not easily be defined or transferred with confidence to non-census personnel, policy interpretation activities, such as those concerned with applying statistical rules and procedures, because non-census personnel could not reasonably make the judgemental decisions involved and the implementation of national policy compliance initiatives, such as preparing briefings for government, preparing the secondary legislation, implementing confidentiality policy and deciding on prosecution in non-compliance cases.

Risks and uncertainties were assessed and threats and levels of risk defined. Two principal threats exist to placing contracts.

- The requirements for the outsourcing activity, are not defined accurately enough and hence the supplier misunderstands what is needed;
- The supplier, although properly briefed, fails to deliver.

Two subsidiary threats were also considered

- the ability to accommodate unexpected changes to census requirements is impeded by the outsourcing arrangements
- the details of the arrangements for outsourcing triggers off adverse public reaction.

The definition of each perceived risk level rating was evaluated and the various options for outsourcing were assessed for risk.

- LOW: Each threat manageable without special precautions
- MEDIUM: Each threat manageable by specialist census office staff
- HIGH: Given the level of Census Office control proposed, threat is NOT preventable with confidence

A strategy was drawn up to manage the risk. The main elements were:

- do not outsource an activity where the perceived risk level is agreed to be high;
- organise effective contracting between the Census Offices and suppliers;
- consolidate effective project management of the overall census activities using the PRINCE project methodology and supported by a strong Project Office focusing on delivering quality;
- arrange to undertake activities early enough so as to permit the execution of contingency plans;

- ensure that there is careful and close monitoring of census outsourcing activities through the use of Time Resource (TR) catalogue entries.

The 1997 and 1999 Census Test will be used to evaluate the proposals and final decisions will be taken as plans for 2001 develop. Since the original study, it has been agreed to reduce the scope of the recommended approach by cutting back the outsourcing of recruitment and payment of field staff to relate to payment only.

- 1 Patrick Heady, Stephen Smith and Vivienne Avery, 1991 Census Validation Survey, coverage report. OPCS Social Survey Report SS1334. ISBN 0 11 691591 a, HMSO 1994
- 2 Estimating with Confidence, Working Paper 10, Dept of Social Statistics, University of Southampton SO17 1BJ, England
- 3 1991 Census General Report, Great Britain, ISBN 0 11 691616 8, HMSO, 1995
- 4 Final report on AGI/Census of Population Round Table, AGI Publication No 1/95, ISBN 1 874059 19 5, AGI, 1995
- 5 Looking towards the 2001 Census, OPCS Occasional paper 46, ISBN 1 85774 204 4, 1996
- 6 2001 Census Information Paper, The 2001 Census Development Programme - as at April 1995, OPCS, 1995
- 7 Census Newsletter No 34, 13 October 1995, OPCS
- 8 Census Geography in the year 2000, paper to AGI conference 1995, John Puckey, OPCS
- 9 Information Paper on Neural Network Imputation Trial, Jan Thomas, OPCS, 1996
- 10 Draxmont Management Consultancy, Walton on Thames, Surrey, England - unpublished report.

13 March 1996

h:\mar96\arcpaper.amc