



# Census 2001 Review and Evaluation

## Downstream Processing: Evaluation Report

March 2004

Content	Page
Project Objective.....	2
Background.....	2
Methodology .....	3
How Well Did it Work .....	9
Lessons Learnt .....	12

ONS is carrying out a review and evaluation of the 2001 Census in England and Wales which will culminate in a Data Quality report and a General Report being published.

Plans for individual reports on specific aspects of the Census operation and a timetable for release have been published.

Each report is written in isolation and is subject to amendments as processing progresses and further information comes to light.

Reports will be released on the ONS website in the form of a high level Executive Summary and a more detailed Evaluation Report.

---

Census Customer Services  
ONS  
Titchfield  
Fareham  
Hants PO15 5RR

**Telephone:** ++44 (0) 1329 813800  
**Fax:** ++44 (0) 1329 813587  
**Minicom:** ++44 (0) 1329 813669  
**E-mail:** [census.customerservices@ons.gov.uk](mailto:census.customerservices@ons.gov.uk)  
**Website:** [www.statistics.gov.uk/census2001](http://www.statistics.gov.uk/census2001)

# Census 2001 Review and Evaluation

## Project Objective

To plan, coordinate and manage the development and implementation of the processes which followed the receipt of data from Lockheed Martin (LM) through to the creation of a clean database for the production of 2001 Census Outputs. These processes and systems were known generically as ‘downstream processing’.

The systems were designed to:

- check for data validity and consistency and to correct data where necessary;
- impose confidentiality;
- assure quality;
- provide information for the production of statistical outputs; and
- establish adequate security and access mechanisms.

The majority of work and resources went into the first of these, data validation, consistency and correction.

(Note that this report deals with the technical delivery and implementation of these systems. The individual project reports for each component cover the processes within them.)

Pinning down requirements was essential and relied upon the cooperation of a number of other related projects. The design, development and implementation required a mix of business analysts who had a good understanding of census matters and skilled Information Technology (IT) resources.

The operational live running was dependent on the phased development and delivery of the downstream systems, (over which the project had total control) and the coded data from LM. The UK was divided into 112 processing areas called Estimation Areas (EAs). The first planned delivery of EAs from LM was 31 July 2001, with deliverables contracted from then until 31 March 2002. However, the delivery from LM was delayed and apart from a handful of EAs, the bulk of the delivery was between January and May 2002. This effectively halved the available time for downstream processing and dealing with this revised schedule placed significant demands on the project.

## Background

To produce credible and reliable statistics from the 2001 Census it was necessary to manipulate some of the data provided by the capture system, as some responses were not captured correctly and forms were not always completed correctly. The suite of systems developed to handle this manipulation of data came to be known as the ‘Downstream Processing Systems’, at the time the work was being considered for outsourcing. A decision was made, largely on grounds of cost and available in-house expertise, to conduct downstream processing in-house. The project was primarily concerned with:

- collecting, agreeing and signing off requirements for each component of the suite, including Edit and Imputation, Disclosure Control and a system to support Data Quality initiatives;
- designing and specifying the systems to meet requirements;
- developing, testing and implementing the systems to meet requirements; and
- operational running of the systems to timetable.

The initial deadline for the live running was to supply the first tranche of Local Authority (LA) output data by July 2002, with the final delivery of the Output Supply Database, from which all standard output is produced, at the end of March 2003. The first tranche of data was required to enable the publication of Census First Results in August 2002, to support the Standard Spending Assessment (the process of resource allocation from Central to local government) by the Office of the Deputy Prime Minister. Because of the revised schedule for data delivery from LM, the initial deadline was revised to August (to enable publication in September) with the delivery date for the Output Supply database remaining the same.

The project ran from July 1999, when the decision was made to develop the downstream processes in-house, to June 2003 and cost around £7m. During this time the Office for National Statistics (ONS) Information Systems Division underwent a number of changes, including a change of name to Information Manage-

# Census 2001 Review and Evaluation

ment Division. Throughout this report all references to Information Systems, Information Management and Information Technology will be as 'IT'.

This report covers England and Wales only although there were many interfaces with the General Register Office for Scotland (GROS) and the Northern Ireland Statistics and Research Agency (NISRA) that had to be considered within the scope of this work.

## Methodology

### Project Management

At the start of the project, governance of the work was controlled within the overall Census Programme and the project reported to the Census Operations Board (COB). COB served as a useful mechanism for bringing together all project managers from related projects. The Board was disbanded in early 2000 and the project subsequently reported to the Census Programme Board (CPB).

### Staffing

One recommendation from the evaluation of the 1991 Census Processing activities was that all IT staff and the relevant census business staff should be co-located to ease communication and problem solving. The IT teams were set up and line managed by senior IT staff who were located along with their business counterparts.

During this project, ONS appointed the auditors KPMG to carry out an efficiency review within ONS. One of their recommendations was that IT specialists should be consolidated away from the business areas into one IT Division. This recommendation was implemented in April 2000 but the downstream processing project manager continued the business management of IT staff and physically they remained within the business area, although they became line managed by an IT specialist.

Staffing was made up of IT and business analysts, skilled programmers, and administrative staff who ran and monitored progress of live running.

### Coherence

During very early planning for 2001, the need for a role to ensure coherence between all systems and technical environments was identified. A Coherence team was set up which gathered requirements from all projects and converted them into the 2001 Census IT Flow Diagram (affectionately known as 'The Wallpaper'). The team grew from one person in 1998, to 3 people throughout the main development and live running phases. This diagram represented all the flows of information between projects, with a high level view of the processes involved within projects, and was created in 1998 for the Rehearsal. A less unwieldy diagram was produced for the final 2001 operation which detailed the downstream processes and their interfaces.

### Requirements Analysis and Design

In June 1998, the IT Strategy Document for the Census made the following recommendations on systems analysis, design and development:

- the Rapid Application Development (RAD) approach would be used;
- the Technical Design Team would work with the relevant development team to produce an outline design to include technical interfaces; and
- the Technical Design Team would ensure the coherence and integration of all census systems.

The RAD approach was described in internal working guidelines, which were based on a well established and widely used analysis and design methodology DSDM (Dynamic Systems Design Methodology). The main emphasis was on iterative development, although the guidelines allowed for a more traditional approach where iterative development was not suitable.

The IT Strategy also had the following recommendations:

- all data would reside on the Sybase database [Adaptive Server Enterprise (Version 11.5)];
- Powerbuilder (Version 6) development language would be used. Exceptionally, for reasons of efficiency or interfacing, Visual C++ or Visual

# Census 2001 Review and Evaluation

Basic or Access could be used with the agreement of the project manager; and

- Windows NT and PC architecture would be used.

In consultation with the business and IT Managers, the Technical Design Team produced an overall system design covering all census IS systems showing their relationships and dependencies. This formed the basis for the division of the system into discrete components for development. Lower level requirements for these components were gathered at Joint Requirement Planning sessions (JRPs).

From May to August in 1998 JRPs were held with all key users. The outputs from these included a visual representation of the process, its associated interfaces and a list of issues to be resolved. Contributors were asked to sign off requirements. Following this phase of requirement gathering, an overall processing diagram was developed together with detailed descriptions of each process. The first draft of the complete system design was available in November 1998.

Also in November 1998, the Census Programme Board announced that it would look at the feasibility of outsourcing further elements of downstream processing including edit and imputation. During this period resources were directed to supporting this evaluation and no significant in-house development took place.

In March 1999, it was announced that due to the cost and because of the risks of an external supplier misinterpreting complex requirements, downstream processes would be developed in house. This development used the output from the earlier requirement gathering exercise as its starting point.

After 1999 there were significant alterations to census requirements. In all nearly 100 requirement changes were documented. These were evaluated for their impact (i.e. how difficult would they be to implement) and for their significance (i.e. how essential did users consider them).

As well as these changes there were significant new developments consistent with upgrading a rehearsal system to a fully functional live system. There were around 70 separate processes in the downstream operation. Other peripheral processes (e.g. the Data Quality Coding Checks) were also developed. While some continued with a RAD approach (e.g. Data Quality Management System (DQMS), Process Control) others had a more formal development pattern where design was based upon a static user requirement.

The user requirement was not always provided by the statistical area. The requirements for a number of processes such as the Longitudinal Survey (LS) sample and Sample of Anonymised Records (SARs) came from business areas outside census. Other processes were defined by the coherence team and agreed by the statistical area. Later processes were developed by IT teams in consultation with the Output projects.

All downstream processes with the exception of One Number Census (ONC) and Geography processes were run against the SYBASE database. The majority of programs were written in a combination of 'C', Transact SQL and Powerbuilder. This mix of programming environments is at odds with the initial recommendations in the Census IT Guidelines. The prevalence of 'C' is a reflection of the inefficiency of Powerbuilder in large and complicated batch processing.

## Software Development and Testing

Once the System Requirements had been produced, the program development began, and consisted of three stages.

- Program design.
- Program development.
- System testing.

In addition to the development stage, customer testing was carried out by the users.

# Census 2001 Review and Evaluation

---

## Program Design

No standards were set down for the way in which program design and specification should be carried out within this technical environment, so each team decided on the best approach for their work. The size of the different systems varied a lot and also some of the systems were batch processes and some were on-line; both of these factors had an impact on the approach chosen.

The purpose of the program design was to work out the most efficient way to handle the data and to determine how to divide the program into modules (a module being a logically distinct function which is a manageable size to be developed in one go). For some of the systems, it was clear how to approach the development and divide up the functions, either because the system was small or fairly simple, or because each function naturally became a separate process. In these cases, it was not necessary to go through the specific task of producing a program design. The majority of systems had specifications written for each module.

For the Load and Edit and Imputation system (EDIS), due to their size and complexity, it was essential to work out program designs. This was done by the IT Team Leaders responsible for developing the system. Each program was divided into modules; specific processes and any common functions were identified. For EDIS, advice was sought from Sybase Consultants about the most efficient way to access the databases and how to process the data once it had been retrieved.

For these two systems, a diagram of the program design was produced showing the flow through the system and the modules.

## Program Development

The batch systems were developed in a 'traditional' manner. Each module/function within the system was developed separately and tested in isolation. Test plans, test data and expected results were produced by the programmer who was developing the module.

The on-line systems were developed using the RAD approach. This involved developing basic screens, then additional functionality was added as required by the users.

## System Testing

For most systems, this testing was carried out by the IT Programming Team Leader and fully documented.

However, as EDIS was very large and complex, a separate team was set up to system test it. A large amount of resource was allocated to system testing so it could be done thoroughly. Comprehensive test plans and expected results were produced. The testing team ran the tests themselves; any problems were passed back to the programming team for correction and re-testing.

**Version Control.** A decision was made not to buy a formal Version Control system, instead, a manual system was implemented by each of the development teams. This was controlled by making one or two people within each area responsible for controlling the procedures between the development and testing environments and documenting the versions. This worked very successfully.

**Application Operating Manuals (AOMs).** AOMs contained details of how to run the system in production. They gave details of the parameters needed and showed which databases and files are input to and output from the system. AOMs were written for each system and were used by the Process Control team.

**Sign-Off Documents.** Each new version of the System Requirements was officially signed-off. The customer testing was also signed off to say that the users were satisfied that the system delivered the functionality required.

**Quality Assurance (QA) Procedures for Code.** The code produced for each system was quality assured, mainly for efficiency, by the Sybase Consultants.

# Census 2001 Review and Evaluation

## System Architecture and Live Running

A Diagrammatic representation of the System Architecture used can be found at [www.statistics.gov.uk/census2001/pdfs/system\\_arch.pdf](http://www.statistics.gov.uk/census2001/pdfs/system_arch.pdf)

Gathering of requirements, the systems development and technical environments have been dealt with elsewhere in this report. However, it is clear that all these had a major role to play in the success of Live Running. What started out as a fairly simple set of processes with built in back-ups finally became a set of some 70+ major processes, 30+ ancillary processes and 20+ database dumps which had to be run for all 112 Estimation Areas (EAs). A Process Control system was developed to simplify the mechanics of running the different jobs and impose process interdependencies. EAs (which had a population of around 500,000, consisting of whole Local Authority Districts) were used as the basic processing units because they were the units upon which ONC Estimation and Imputation was based. The diagram available from [www.statistics.gov.uk/census2001/pdfs/down\\_process\\_flow.pdf](http://www.statistics.gov.uk/census2001/pdfs/down_process_flow.pdf) gives a fairly high level view of the main processes and interfaces involved.

A detailed timetable was developed in conjunction with a procurement exercise for the processing and image servers to enable ONS to come up with the best technical environment for the processing required.

The known processes, processing order and the comparative predicted run times for each process, along with the degree of processing intensity required for each of the processes, were listed with assumptions about which processes could be run concurrently on a single server. A volumetric exercise to predict the size of all databases, input files, output files and error logs was based on information from volume testing and Rehearsal. The processing order was based on the contracted delivery timetable, assuming a smooth delivery of EAs from start to finish. This was translated into expected completion dates for each process for each EA, including allowances for all manual and external processes. The timetable and hardware had sufficient contingency and additional capacity to cover minor fluctuations in delivery.

Processing times varied, from a few seconds to 10 days. Most came within the 1-3 hour bracket with a few, which involved donor searches, taking between 3 hours and 10 days. Because of the volume testing which had been carried out, and the early EAs from Lockheed Mar-

tin (LM), we were able to predict the problem processes and tune them as far as possible. This meant that few processing times came as a surprise.

LM, who scanned the forms, provided their images and data on Digital Linear Tape (DLT) 7000 tapes in native NT format. Deliveries were received at ONS by the Throughput Team, who, after checking the delivery, passed it to the Process Control team. They also checked what they were receiving then signed a receipt form.

Due to the enormous numbers of tape movements required to ensure the data and images were restored and the tapes archived, a database was devised to ensure electronic movement records of all the DLTs. This worked extremely well and saved a great deal of time.

A good dialogue was established between LM and ONS staff at Widnes to establish what would be delivered by when. This enabled the downstream teams to plan work efficiently. LM moved from one large to two smaller deliveries a week towards the end of processing. This worked very well from a processing point of view as we were able to manage the processing flow more effectively.

The average time for Images of census forms to be restored to the ONS servers was anywhere between 6 and 12 hours. The images had to be restored in optimum time in order that any queries with the data could be resolved by interrogating the images. The data was restored more quickly, on average 30 minutes per EA. The processing to get the image indexes loaded onto the system checked both copies to ensure they held the same information and re-formatted the data before bulk copying onto the database by the Sybase Support Team.

Procurement of the hardware to support the image management system was linked to provision of the hardware used for downstream processing. Within this contract there was scope to increase the quantity of servers and storage purchased without the need for a further procurement exercise. This also ensured that we could standardise the equipment that we would use as far as possible. The system proved robust and although there were occasional server failures these did not affect processing significantly.

# Census 2001 Review and Evaluation

Once the data and images were restored, the Load process could be run. Because of sporadic deliveries during the early months, there was time for the Coherence Team and a small team of casual workers to carry out a lot of data quality checking. This led to some improvements being made to the Load processes.

The work also brought to light other quality issues, in many cases caused by enumerators' poor handwriting, which resulted in invalid geographical identifiers being attached to records. These could be corrected by looking at address information. In addition, there was some poor enumeration of Communal Establishments, resulting in students or residents in rest homes being incorrectly linked to households by the LM systems. Ad hoc checks were developed to detect the problems and these were run by the Coherence Team before the Load process. Analysis and correction of problems flagged was a fairly laborious manual process, also carried out to an extremely tight timescale (usually within 1 day of delivery) by the Coherence Team.

Further problems were flagged by the Load Process which had to be analysed and corrected by the Coherence Team. These were mainly aimed at reducing the number of households and communal establishments rejected because they were not found on the Census Geography Database. As soon as the EA had completed the Load processes, the need for manual intervention reduced significantly. The loaded data and images were used by ONC for matching with the Census Coverage Survey (CCS) data, while the data went on through the Edit & Imputation process. These two activities took about a week per EA, including time spent by the EDIS/Quality teams to provide data file amendments for any households containing missing data which had no suitable donor.

Once the Edit & Imputation processing was complete, data was extracted to feed into the ONC Estimation and Imputation systems, selected on the basis of the ONC Matching exercise. Downstream Processing could not continue until ONC Imputation was complete and Quality Assured. In the early days of sporadic deliveries this was completed to the planned timetable, but when we started receiving 6+ EAs per week, the ONC QA process caused a large bottleneck which in turn caused significant delays. In addition there were some problems with the ONC Imputation system which meant all EAs had to be rerun.

In August 2002, it became apparent that the ONC figures needed an adjustment for dependency, which meant all EAs had to be rerun from the ONC Estimation and Imputation stage. This required some late changes to the processing schedule to streamline the work to ensure it would meet the new deadlines. It also meant that database dumps for all EAs had to be restored, and the Process Control team had to produce a new set of documentation for each EA to record the day to day running required for each process to run. The processing servers were fully utilised for a six week period with overnight running where possible and weekend working. This inevitably involved some chopping and changing and swift workarounds to maintain the planned throughput. Without the cooperation and goodwill of the Database Administrator & Support team, and the Process Control and Coherence teams, this work would not have been completed to this challenging timetable.

Downstream Processes beyond ONC Estimation and Imputation comprised an Update to impute the households and people identified as missing by ONC, followed by a further pass of Edit & Imputation to ensure their relationship information was complete. This was followed by Disclosure Control, further imputation of missing or incomplete migration and workplace postcodes, the Household Composition algorithm and various processes that were developed in response to problems found during processing by LM, and subsequent downstream processing which included:

- a process that enabled the addition of population from late delivered forms;
- three processes that allowed identification of areas that were affected by black lines on images and adjust the data to compensate for the incorrect responses generated; and
- a 'final consistency check' to reapply the filters used during the Load process and flag various discrepancies.

Following this, the Output Supply database was created, followed by a host of processes that prepared the data for outputs. In total some 15,000 processes were run within the Downstream Processing system, including 1,141 Data File Amendments (DFAs).

# Census 2001 Review and Evaluation

## Issue Management

All issues were raised either through the downstream processing database or through the more formal Data Query Resolution Process, details of which can be found in the Quality Report (due to be published Spring 2004).

## Security

An independent Census 2001 Security Review was undertaken by HEDRA in October 2000. This covered the full scope of 2001 Census activities including field, external processes, the three Census Offices and the links between them.

Within downstream processing at Titchfield the ONS security standards were used for NT server build and NT workstation access. Data and Application security was provided by NT and/or Sybase security.

Remote access to Downstream Processing by NISRA and GROS was provided by a dedicated 'Census WAN'. Global Crossing provided the Government Data Network (GDN) LAN interconnect service linking ONS, NISRA and GROS together with the main processing site. Census data is protectively marked as Restricted although the Data Custodians additionally required encryption of data across the WAN. This encryption was originally to be provided as part of this service for Global Crossing but technical issues meant this was actually provided using in-house encryptors.

A Lotus Notes workflow system was used to authorise and record access to data and applications.

## Other Related Responsibilities

This project was also responsible for providing other areas of ONS with 2001 Census data.

- The **Longitudinal Study** for England and Wales follows a one per cent sample of people throughout their lives using census and life events data. Data from each census are added to the data repository.

- The **Sample of Anonymised Records** (SARs) which provides a Household and Individual extract, from the 2001 Census, for supply to the Cathy Marsh Centre (CMC) at Manchester University. Household and Individual SARs were first supplied following the 1991 Census. For 2001, CMC requested Small Area Microdata (SAM) in addition to the Household and Individual SAR.
- The **2001 Census-linked Survey of Non-response** being carried out by Social Survey Division (SSD) to examine the characteristics of responding and non-responding households. Census provided SSD with a data-set for those census households, and associated persons, which matched a list of survey households supplied.
- The **Social Exclusion** sample requested by the unit looking at Area Classifications.

## System Interfaces

Each interface with other systems was treated individually, with a general rule to keep things as simple as possible. A decision was taken early on to have a set of tab delimited text files delivered on Digital Linear tape (DLT) in native NT format, one for each of the database tables that would be on the Census database. An intrinsic part of the 'technical' interface was the need to interface with the people running the process at Widnes. Formal interfaces were developed between the Process Control throughput teams to ensure there was a record of what was delivered, and how any tapes were to be returned. ONS staff at Widnes reported many problems that would manifest themselves downstream, and we had an opportunity to be involved in decisions on problem resolution.

It was also decided, at an early stage, that there should be a single source of Geographic data. The contract for the procurement of a Geographic Information System (GIS) was let, with one of the mandatory requirements being, a seamless connection to the RDBMS and development software for 2001 Census processing. Sybase offered a direct connection to RDBMS (Oracle) which, once set up, proved to be robust.

The technical interface between ONC and the main-stream downstream processing was developed quite late. It consisted of comma delimited text files being passed

# Census 2001 Review and Evaluation

---

between the various processes rather than direct database connections. The ONC Matching process used an interim copy of the Census and CCS databases on a separate server for matching while the mainstream processes continued to update the 'real' databases. The subsequent processes within ONC were written using SAS which was not able to access the Sybase RDBMS directly, so the sequential file interface was the obvious option.

The dependencies between the ONC and downstream processes were such that the data had to be processed through EDIS before ONC could run their Estimation, and the ONC Imputation had to be complete before downstream processing could proceed beyond EDIS. Once the main census database had been updated with the additional people and households identified by ONC as missing, downstream processing had to pause until the data had been Quality Assured.

SuperCROSS software was chosen to tabulate census outputs which did not interface directly with any RDBMS. Therefore, the chosen method of interfacing was a tab delimited text file for each EA which are loaded into a 'Country' based SuperMART database.

There were nominated single points of contact, in GROS and NISRA, for decisions concerning the data for their country, and for input to any issues or requests for change. They also took responsibility for the quality of their data at the Load stage, and edited the data where necessary.

## Output Design

The output database was designed independently. The requirements for output did not begin to evolve until consultation with external and internal users was completed in 2001. The output database reflected the requirement to tabulate data and contained fields and relationships which could have compromised the efficiency of the input database.

Output database design used standard SSADM techniques and the tool used to document and generate the design was POWERDESIGNER DATA ARCHITECT marketed by Sybase. POWERDESIGNER was integral to the success of the work since it simplified the design process.

It was important for downstream processing that a single design would work for all of the UK. This removed the overhead of developing country dependent processing. The downside of this approach was that the database has fields which are redundant in some situations and the three census offices had to reach a compromise on the final schema.

Requirements were incrementally enhanced through a continual QA process beginning with the first release in June 2001. This enhancement process ended in September 2002 with the seventh and final design. Intervening releases were used by programming teams as they became available.

## Process Control

A process control system was written to enable ONS to run all 112 EAs, plus their backup procedures and subsidiary tasks in a controlled and managed way. Once the system had been developed and tested it was handed to a clerical team to implement and manage.

## How Well Did It Work

### Project Management

Generally, the division of all activities into projects worked well.

### Staffing

There were a number of difficulties in retaining skilled resources for the duration of this project. The solution was to recruit contract developers and to award a retention bonus to the in-house team if they remained with the project and met certain performance criteria. ONS also procured additional consultancy to assist the developers with training and understanding the new technology, and to provide a quality assurance role to ensure design and coding were efficient. Overall this combination of a mix of in-house and contract staff worked well and contributed to the overall success of the development cycle.

# Census 2001 Review and Evaluation

---

## Coherence

The setting up of a coherence team was a key factor in the success of this project. Use of the RAD methodology which enabled staff, from projects which interfaced, to get together through Joint Requirement Planning (JRP) sessions and to agree those interfaces, was generally well received. A good rapport was built up during the JRP sessions that enabled cooperation and 'buy' in from all census projects.

Work that went into the production of the flow diagram enabled the coherence team to retain an overall picture of where things were going and what needed to be done to get there. This knowledge allowed the team to play a pivotal role in keeping downstream processing on track and ensuring the quality and coherence of the data as far as possible. In addition, it raised the awareness of the various policies that underpinned the statistical and quality systems, enabling ONS to ensure that these policies were not departed from.

This also provided the coherence team with sufficient information to allow detailed live running timetables to be developed, and the operational environment to be defined. This information also allowed quick investigation of contingency measures whenever it became clear the timetable may need amending and provided detailed information to underpin management reporting.

Consistently 'keeping it simple' meant that we always had a sound platform for any rerunning, using both the database dumps taken at every update stage, and any input files, the database could always be recreated at whatever stage of processing was required. No system developed by anyone other than the Downstream Processing Development teams could update or extract information from the database. This meant the integrity of the data was reliable, and there was full control over any data entering or leaving the systems. This also meant that ONS could ensure the database, images, and ancillary data were subject to strict security and access policies, and there was a record of all access permission's granted.

## Requirements Analysis and Design

The approach to requirements analysis worked well for many of the census processes. The early JRP sessions and the subsequent Processing Flow Diagram formed a

sound bedrock for subsequent analysis. There was also a strong element of prioritisation in processes such as the Edit and Imputation System (EDIS) which echoed the 'Must have', 'Should have', 'Could have' precedence laid down in RAD guidelines.

## Software Development and Testing

Use of experienced in-house resources in planning, estimating and delivering complex systems was key to the success of this project. Many of the team had previous census experience and this 'topic knowledge' should not be under-estimated in any future census development activities. Despite the new software environment and the new skills the team had to acquire, all development was completed to timetable.

## System Architecture and Live Running

There were few problems during production running because the systems architecture was fully tested. Planning and installation of servers had included sufficient time to iron out any teething problems.

The procurement contracts included our contingency plan for increasing the number of servers. We were therefore able to procure additional servers and storage quickly when it became clear we would be getting a higher rate of delivery from LM than originally planned.

The volume testing carried out was invaluable as it highlighted the need for tuning in several processes, and the need to split some of the larger EAs down into smaller groups for the EDIS to reduce run-time.

The database devised for movement of computer tapes worked extremely well and overall the paperwork to handle receipt and rejection of data and image tapes proved practicable and the interface with the various teams worked. Downstream processing staff suggested the purchase of sets of 'turtle boxes' for transfer of the tapes between sites. This eliminated the need for acclimatisation so tapes could be loaded as soon as they were received. The software handled the restoring of the large volumes of images with very few exceptions.

# Census 2001 Review and Evaluation

---

The Process Control system was flexible, and able to skip processes in the schedule or move them around if required. Management of the database and production environment was handled efficiently by both the IT and Process Control team, with all involved being helpful and responsive. All 112 databases were kept on-line, with database dumps from key points also available to both the Data Quality Management System (DQMS) environment for quality checks and the live environment. This allowed the minor and major rerunning that took place. Restores never took long. Large jobs were restartable mid way through a process. On-line dumps were actively managed to ensure we did not run out of room, and to maintain a sufficiently comprehensive set of dumps for use by people investigating the data.

Most automated processes ran in less time than estimated. Having a detailed timetable allowed the anticipation of problems with delivery and come up with solutions before there was any serious delay. The timetable was able to cope with changed requirements (e.g. deriving ONC variables after Edit & Imputation), new requirements (addition of 'missing forms' where LM found forms that had 'not been delivered') and unexpected problems (for example 'correction' of problems caused by black lines on images).

The fact that we had planned for Data File Amendments (DFAs) in advance and developed a small system that applied range checks and audited the changes proved very important. All changes applied to the data using the DFA process were audited to the same specification as all other Downstream systems. This proved invaluable when checking what changes had been applied to the data.

In spite of an original directive that only LS and ONC would require on-line access to the census form images, and therefore only 10 per cent of images would be required on-line at any one time, we ultimately managed to hold images for all 112 EAs on-line. This was in response to the number of different areas needing access to images at different points in the operation. Originally it was planned to hold images on-line for 6 weeks, but managing the removal and replacement of images proved to be impossible because of the differing needs of image users. Images have provided a useful back-up to microfilm at present, with 2.2 million forms

being filmed from image. Use of images means that we no longer have to provide storage for around 25 million forms, a significant saving for this census.

Good team-working and communication was maintained throughout the processing, even during times of extreme pressure. This was mainly down to the appreciation of the importance of each other's roles and the high level of commitment each team member was prepared to make.

A key to the success of Live Running was the flexibility of people working on all aspects of the processing, who were able to come up with quick, well thought through solutions to minimise extra work. In addition, the flexibility of the timetable, systems and technical environment were also major factors contributing to its success.

## System Interfaces

During development and specification of the processing requirement, staff worked together to ensure everyone was clear on the requirement from Lockheed Martin (LM), and how we wanted them to process the data. LM generally adhered to the agreed formats set out in the Interface document, although the Load process was changed to accept a few minor deviations from the specification.

The LM Audit file provided the only way to establish exactly how a form passed through the LM system. The development of the LM Audit file took considerable cooperation, between downstream coherence staff and LM staff, to reach a mutual understanding.

Three sets of tests, Operational Readiness Test (ORT), System Functionality Test (SFT) and Total Operations And System Test (TOAST), carried out between LM and ONS, provided a valuable opportunity to test the Load process and adjust both our expectations of what LM would provide and correct their systems. Communications with LM worked well and records of responses and decisions were documented.

The Geography team consistently delivered Geographic processing required to timetable. The physical link between the downstream processes running against Sybase and the Geography database held on Oracle was robust.

# Census 2001 Review and Evaluation

---

The way the Geography database was designed allowed each UK country to look after its own geographical data sets.

## Output Design

The use of POWERDESIGNER was crucial to streamlining development and deployment of the design and made maintenance and change management simpler. It supported database structure generation, documentation, and reverse engineering. Having a centralised design team allowed simple version control which was essential given the iterative nature of the design process.

The decision to resist turning the output database into a statistical evaluation and analysis repository was sensible. The focus on the output database as a source for tabulations was correct.

Changing requirements were a characteristic of census development. Within this context the output database design process was successful because it was flexible and because the design tool was effective.

## Process Control

Process Control was set up with good lines of communications which proved invaluable during the rerunning when everyone was trying to keep to the tight timetable. Much time and effort went into liaising with various teams in order to understand where Process Control fitted into the larger picture. Once information had been gathered and understood, clear and concise instructions were produced, tested by doing dummy runs and refined where necessary.

Forms were designed to control the movement of tapes and CDs and ensure an audit trail for the delivery of tapes/CDs was available for ONS and LM. Electronic diaries/spreadsheets allowed senior managers to keep in touch with progress and also provided an audit trail for every EA. Labels were produced for all EAs, ahead of their delivery, to help with the smooth running of deliveries and this saved valuable processing time when deliveries were made.

Process Control ensured that when a new program was released the relevant developer produced an Application Operations Manual (AOM) detailing the expected output. These clearly documented any errors that may arise during the running of a process and procedures to be followed, and were a very important tool used by the team.

The team met all their deadlines largely due to the staff within the team being flexible, project focused and having the ability to build good relationships with all downstream areas. The team were willing to stay late, change their working patterns or even change their leave as necessary.

On the whole the Process Control system functioned well and provided ONS with the tool to carry out the processing of a very large data set. The control system was also accessible to nominated GROS and NISRA staff. This proved beneficial to the overall running of the processes.

## Lessons Learnt

### Project Management

There were weaknesses in the overall Board structure especially after reorganisation which led to some difficulties determining who had responsibility for taking decisions. Therefore Downstream Processing often had to work on assumptions and informal decisions rather than clear-cut decisions, resulting in an element of risk and uncertainty. In a future Board structure it should be made clear to all what the mechanism is for making timely decisions. This could be supported by a more robust Programme Management set-up with the expertise to understand and advise on interdependencies and impacts across the whole programme.

### Staffing

As mentioned earlier, ONS centralised the provision of IT resource during this project. From the perspective of a particularly large programme such as the Census, it will always be necessary to have dedicated IT resource and a considerable degree of direct control and management. For ONS as a whole, however, there are

# Census 2001 Review and Evaluation

---

advantages in having IT staff centrally managed. For the future, much will depend on the way ONS as a whole is organised. An important factor in the success of this project was that it had the priority and funding to pay for additional skilled resources as required. In order to retain good, appropriately skilled staff a bonus was offered to those in-house staff who stayed committed to the project. This was hotly debated at the time, but none the less worked well and proved not to be divisive as some had feared.

## Coherence

Although some problems, encountered along the way, were discussed at JRP sessions, and some solutions thought through at the time, not all of these were implemented, mainly because the risk was considered to be small, and there was too much else to do at the time.

During delivery from LM the coherence team were under a great deal of pressure to edit the data to correct enumeration & recognition problems within a day of receipt, as well as being relied on for advice and support for other processes later in the processing order for EAs already delivered. In addition to this, the timetable had to be monitored and management reports had to be produced, as well as liaising with Geography and ONC to ensure that backlogs did not build up. This meant that the team had no leave for a long period, and worked long, very concentrated, hours to ensure that the data was of good quality and the timetable was adhered to. This level of commitment (also applicable for many other areas within the Census operation) should not be assumed in any future planning exercise.

## Requirements Analysis and Design

A realistic assessment of resources required for such a project is needed early on in the project cycle. The requirements for some systems were too complex given the level of resource and time available. For example, the imputation system within the Edit and Imputation part of Downstream Processing was far more complex than the imputation system developed within the ONC system. However, it could be argued that the overall impact of a simple imputation system within ONC had more impact on the overall quality in terms of missingness i.e. missing people versus missing data items. Ensuring that all requirements gathering and resulting design is

managed and coordinated by the same design team may lead to a better understanding of the potential issues and their impact.

Requirements and decisions need to be agreed and signed off at an earlier stage. The late decision on what would be outsourced and what would be delivered in-house caused severe time constraints. Additional funding and resource had to be acquired to overcome these.

The design work carried out within this project did not include data collection and for the future we recommend that it should. Form content, including form identity, form control and geography interfaces were fundamental to the processing system but the project had little involvement with their design. For example the form identity code was a series of alphas and numerics which had to be hand written by Enumerators on each form. This data item was fundamental to a number of systems and bad hand writing along with mis-recognition at the scanning stage caused processing problems further down the line. A coordinated approach across all projects needing to use a common identity code may have prevented this.

## Software Development and Testing

Although the original requirements for the edit and imputation system were reduced, the system was still complex and difficult to develop, and therefore test effectively. The bulk of the testing was carried out by using IT resource. Decisions about the System Requirements must be made by users much earlier than they did for 2001. Too often, development had to start without a clear idea of what was needed. In addition, many requirements changed constantly and were allowed to continue changing throughout the entire development stage.

It is essential to have continuity within the customer area, so that knowledge of the system is retained. Several of the systems had a change of customer personnel part way through the development which caused a lot of problems for the IT programming teams.

# Census 2001 Review and Evaluation

---

Some of the customers had not previously been involved in the complete development life cycle of a system. This meant they did not understand all the different stages and the importance of their role within them. This was particularly true of the customer testing role, which was not always taken as seriously as it should have done. The thought was, that as the programs had been system tested, that should be sufficient. There was little understanding of the concept that they needed to test the system to ensure the functionality fully met their requirements.

In a future census, it will be important to ensure adequate customer testing resource is available in business areas and to ensure that those with responsibility for this important role are familiar with the concepts of version control and rigorous testing needed for large scale production systems. There would also be advantages to including all development activities within a downstream processing project, or its equivalent. Building and implementing large computer processes to effectively manage, control and process this amount of data needs a strong IT input/background.

Accountability and ownership for these processes needs to be with the project manager responsible for delivering the systems and not the people responsible for the methodology.

## System Architecture and Live Running

The Storage Area Network (SAN) technology was fairly new within ONS and the lack of experience in-house and from regular engineers meant that when problems occurred resolution was often painfully slow. As SAN components were not mainstream there were shortages of certain components and alternatives were difficult to acquire. The long lead time for procurement, delivery and installation was frustrating. Eventually ONS bought in spare equipment and components so they were readily available on site. This needs to be considered in any future exercise.

If a similar processing strategy is implemented for the next census, the plan should be to keep everything on-line at all times (cost of hardware and storage restricted this option but with continued reductions in storage cost this may well seem obvious in the future).

There were a number of serious problems with processing the data from the point of receipt from LM which were largely due to changes to the delivery timetables. The delivery pattern should have been contracted and controlled more tightly. The initial processing timetable was based on an even delivery pattern from 31st July 2001 until 31st March 2002. This did not happen. The bulk of deliveries came between January and early May 2002, which needed an increase to the number of servers required to cover the increased processing load. Due to the good relationship between LM, the processing team at ONS, and the downstream team, we had sufficient advance warning of the change to the delivery schedule to allow the lead time to acquire and install additional servers as allowed for within our existing contract.

The deliveries from LM were originally planned to be once weekly, with three EAs expected per week. In reality, delivery often happened three times a week with numerous EAs arriving at a time. Arrival at ONS (Titchfield) was at times late in the day which meant negotiation with the Computer room to stay to take delivery of tapes into their safe for safekeeping overnight. This delivery routine was less than ideal and led to very late working in order to get the data restored on the same day as delivery.

LM were unable to deliver to the agreed EA order which had a knock-on effect downstream because the order had been devised to ensure that adjoining EAs were delivered at the same time, allowing the ONC Matching process to be completed for each EA as quickly as possible. The nature of the dependencies between the ONC processes and the downstream processes was such that an EA would complete Edit and Imputation and then get no further until Matching was complete, therefore this caused further timetabling problems.

For a very few EAs where a significant amount of data was found within the LM system after the EAs were supplied, a 'delta' delivery was organised for that data. A method of inserting the data was devised downstream so we did not have to start processing again from scratch.

Bunching of deliveries caused scheduling difficulties for the Geography staff who had to deal with problems with incorrect geographic identifiers. This also had a knock-

# Census 2001 Review and Evaluation

on effect to downstream operations because processing was dependent on the completion of any postcode changes resulting from correction of these problems.

Because of unforeseen problems with enumeration, data from LM had to be checked and edited before it could complete the Load process. To maintain the processing timetable, this work had to be completed within a very tight time scale. This meant that the people doing the editing had less time to fulfil their coherence role, that is, knowing what was going on, and planning for future processes. This impacted on the organisation of some of the later processes and meant that we were not as clear of interfaces and processes as we had been for the early processes.

Although there was a formal route for data quality problems, it was not clear how problems with the data from LM, flagged during the Load stages, were to be dealt with within the 'Quality Management' umbrella. Initially many decisions remained outstanding. This resulted in the Coherence team making many pragmatic decisions as to how data from LM could be altered to ensure acceptance by the Load process, allowing processing to move on. The line between Data Quality and Coherence became very blurred.

The first planned set of counts to be produced from the data were the age and sex counts required by Population Estimates Unit (PEU). Because of the late requirement to introduce further disclosure control methods and the need to adjust some of the geographic identifiers on individual records, the method for producing these counts had to be changed, and the order for running the individual processes had to be amended.

The 'Data File Amendment' (DFA) process which was developed to allow changes to the data to be applied relatively easily was over used. The original premise was that it would only be used occasionally by people who knew all the data dependencies to correct problems we knew would occur. For example the postcoded fields - enumeration, workplace and address 1 year ago. However, GROS and NISRA used this facility to make many changes to the data. It proved time consuming to check these thoroughly since they affected so many records, and has been difficult to maintain useful documentation as to the variables the DFA was intended to change and why they needed to be changed.

## In summary

- Don't make assumptions about delivery timetables from a contractor - ensure the agreed delivery pattern is tightly adhered to by the contractor.
- Make it clear at the start (or as early as possible) what constitutes acceptable and unacceptable data quality.
- Keep the live running and processing schedule simple; it allows for flexibility when things don't go as planned.
- Ensure any issue management or query resolution system works quickly and effectively so as not to hold up vital processing time.

## Security

The independent 'Census 2001 Security Review' carried out by HEDRA in October 2000 required resource from census IT and business areas. Any future project should plan sufficient time for an independent security review (if required) and implementation of any particular Data Custodian requirements.

Global Crossing provided the Government Data Network (GDN) LAN interconnect service linking ONS, NISRA and GROS together with the main processing site but despite an early request for them to provide encryption as part of the service, they were unable to deliver on time and this part of the service was actually provided using in-house encryptors. In future more time should be allowed for commissioning external services.

## System Interfaces

### Lockheed Martin (LM)

The export facility within the LM system that pulled together all the information for each household and its people proved to be more complex than LM had realised, and took longer to run than expected. The process prevented export if any linking problems were found, these caused such problems for LM delivery that eventually the ONS Quality Manager had to make decisions on what data could be forced through the system, and what had to be rejected and consequently lost for each

# Census 2001 Review and Evaluation

---

EA. There was always an expectation that the in-house systems could be changed to correct known problems (e.g. 3 systems to reduce the effects of black lines on specific data items) rather than asking LM to correct or enhance their systems. This needs to be considered in any resource planning for future exercises.

## Geography

The design work carried out in the downstream processing project did not encompass data collection and for the future we recommend that it should. Form content, including form identity, form control and geography interfaces were fundamental to the processing system but the project had little involvement with the design of those. For example the form identity code was a series of alphas and numerics which had to be hand written by the enumerators on each form. This data item was fundamental to a number of systems. A coordinated approach across all projects needing to use a common identity code would have prevented this.

## ONC

The nature of the ONC meant that considerable manual intervention was required to run the systems. For this reason, the ONC team were responsible for running their area of processing. Much effort was dedicated to integrating the ONC processes into the rest of downstream processing. In the main, this integration and the close working relationship between the two teams worked well. However, there were a number of occasions where lack of experience within the ONC team led to issues, particularly with version control. This was not helped by the delays from LM, which meant that some EAs had to be stockpiled at the QA stage, which in turn affected the smooth throughput required for the rest of the processing. In future, consideration should be given to ensuring the responsibility for production running of all the processes resides in one area.

## Outputs

Providing the required data extracts to outputs in a suitable format to cross tabulate with geography data was not a simple task and the expectations of how well the SuperCROSS software could be used were too high. Initially there were no formal mechanisms in place for feeding back issues between downstream production and output production. This evolved once extracts began to be delivered. A more formal communication

process on data quality issues and decisions was needed. There was no central record of all the issues that had been raised or their resolution, resulting in a lack of clarity about decisions that had or had not been made.

## GROS and NISRA

GROS and NISRA, having a smaller volume of data to handle, made more manual changes to the data, in addition to the 'automatic' processing route. A large amount of data editing took place on the prime data (delivered by LM) on the basis of the error logs produced by a first run through the Load process. This meant there was additional rerunning (as it tended to be an iterative process) and many subsequent Data File Amendments. Checking and keeping track of the DFAs became almost a full time job for one member of the Coherence team for quite a long period. The result of this was more consistent data for Scotland and Northern Ireland, but given the available resources and the volume of data, such an approach was not feasible for England and Wales.

## Output Design

The large number of releases of the design reflects the fact that the full requirement was unclear at the outset. The design also had to compensate for the limitations of the SuperCROSS tabulation package in handling complicated derivations. These limitations were also unclear at the outset. This was particularly true of migration and travel. Since GROS had greater experience of the tabulation package they took on the task of analysing the requirement from a SuperCROSS perspective.

The output systems design work did not start until well into 2001, so systems had been developed in Output product order as far as possible. Delays to decisions on the way Migration and Transport variables would be derived resulted in some delays in the specification and development of these variables.

## Process Control

Although no major problems occurred which prevented processes being run to timetable, there were a number of incidences of servers crashing. On occasion there was no support cover, particularly early in the mornings when the Process Control team required technical input be-

# Census 2001 Review and Evaluation

fore they could move jobs on. In any future exercise the need to enable the control team to do certain database activities needs to be considered.

Service Level Agreements should be set up with the various areas within downstream processing, technical support and the computer room to ensure there are no misunderstandings.

Good forward planning of the procedures/paperwork to be used in the area is essential, along with good communications and relationships with other areas.

## Summary of Lessons Learnt/ Recommendations

- There were difficulties with the matrix management approach used in the project. For future exercises it will be necessary to ensure that there are adequate resource and commitment from internal suppliers to deliver. This will be even more important with the current, ONS project organisational structure.
- Programme Board structure needs to reflect operational working so that appropriate decision making processes are in place.
- Programme Management support staff should ideally have the expertise to understand project interdependencies so that they are better placed to understand the impact of new proposals/changes.
- The project benefitted from retaining a group of skilled IT staff throughout the duration of the project cycle, leading to greater business continuity.
- The coherence role was vital to the smooth operation of the production cycle. Consideration needs to be given to increasing the size of this team, in any future similar set up.
- Requirements need to be agreed and signed-off earlier.
- Ensure adequate customer testing resources are available in relevant business areas.
- All IT equivalent activities should be controlled and managed from within the Downstream Processing Project, or its equivalent.

- Accountability for these processes needs to be the project manager responsible for delivering the systems and not the people responsible for the methodology.
- Use a scalable design for processing architecture to allow for ramping up of processing power and/or storage capacity.
- Allow a long lead time for procurement, delivery and installation - then allow some more time.
- Regard new technology with healthy suspicion - some of the tried and tested technology was necessary to supplement certain SAN elements.
- Plan to have everything (data and images) online.
- Don't make assumptions about delivery timetables from a contractor - ensure the agreed delivery pattern is tightly adhered to by the contractor.
- Make it clear at the start (or as early as possible) what constitutes acceptable and unacceptable data quality.
- Keep the live running and processing schedule simple; it allows for flexibility when things don't go as planned.
- Ensure any issue management or query resolution system works quickly and effectively so as not to hold up vital processing time.
- Allow plenty of time and more for the commissioning of external services such as a Government Data Network (GDN) LAN to interconnect the various sites.
- Plan sufficient time for independent security review (if required) and implementation of any particular Data Custodian requirements.
- Some processes and Data File Amendments affected the consistency of the data. In future we should plan to develop a final detailed integrity check, having mapped all the data dependencies at the outset.
- Ensure sufficient resources available to 'fix' unexpected quirks/omissions/errors in the delivered data.
- Plan to include data collection activities in the overall design to avoid the un-coordinated approach to the design of the form identity code.

# Census 2001 Review and Evaluation

Census Topics	Target Dates for Release
Legislation	Published
Non-Compliance (Executive Summary Only)	Published
Data Needs	Published
Geography	Published (Executive Summary)
Publicity	Published
Data Collection Development	Published
Data Collection Support	Published
Census Coverage Survey	Published
Processing	Published
Annex: Quality of Data Capture and Coding	Published
Downstream Processing	Published (Executive Summary)
Data Quality	
- Question non-response rates	Published
- Disclosure Control (Executive Summary only)	Published
- Data Validation (Executive Summary only)	Published
Edit & Imputation	Published
One Number Census	
- Quality Assurance	Published
- Lessons learnt (Executive Summary only)	Published
Output Policy	Published (Executive Summary)
Output Production	
- Part 1:Review of Output Released to date	Published (Executive Summary)
- Part 2:including Sample of Anonymised Records (SARs)/Origin Destination Matrices	Published
Census Access	Published
Programme Management	Published (Executive Summary)
Quality Report	Published
General Report	Published

Please note that the dates for release of individual evaluation reports noted above are target dates, and therefore subject to change. For the latest information please visit [www.statistics.gov.uk/census2001/reviewevaluation.asp](http://www.statistics.gov.uk/census2001/reviewevaluation.asp)