

Census 2001 Review and Evaluation

September 2003

Quality of Data Capture and Coding: Executive Summary

Content	Page
Background.....	2
Methodology	2
The ONS Checks on Data Quality	3
Results.....	4
Conclusions	5

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
Website: www.statistics.gov.uk/census2001

Census 2001 Review and Evaluation

Background

For the 2001 Census, the task of setting up the Processing Operation to extract data from the census forms and code text responses was contracted to Lockheed Martin (LM). LM decided to base their system design on the Office for National Statistics (ONS) model for 1997, using scanners, optical mark and optical character recognition, followed by automatic and manual coding for text responses.

The Processing Operation was to capture and code the data from 27.3 million forms, representing 59 million people.

There were 18.453 billion tick boxes and 6.099 billion characters captured from the forms, and of these, 207 million tick boxes and 1.057 billion characters were sent to operators for correction.

Once captured, written responses to questions were coded to standard classifications to enable analysis. The total number of text responses coded was 94.68million, with 68.04million of these coded automatically.

Central to this huge data capture exercise was confidence in the quality of the output data. Part of the contract with LM was a requirement that they would measure, monitor and report on the quality of data. Service Level Agreements (SLAs) for each type of capture and coding were set within the contract. LM produced a Data Quality Management Plan (DQMP) that detailed how they would measure and report the accuracy of their processes, identify systematic error, and the steps they would take if accuracy fell below the agreed Service Levels.

It must be noted that this report is an account of the history of the Quality Assurance processes and analyses. The accuracy reported for data capture was at an early stage in processing, and corrections were made during some of the following processes, so this report does not directly reflect the quality of the data that is output. Further improvements in accuracy were made and these are reflected in the published data.

Methodology

The Lockheed Martin Quality Assurance System

The Data Quality Management Plan (DQMP) described the systems and procedures LM would put in place to measure and report accuracy and the steps they would take if accuracy fell below the Service Levels, or if they identified systematic error in the data.

LM incorporated an automatic Quality Assurance (QA) system within the data capture and coding subsystems. ONS took the lead in testing the QA system to ensure that the counts were being taken and reported in line with the DQMP. These tests isolated faults in the QA system, and were iterative until June 2001, by when all major faults had been corrected.

The DQMP detailed the sample sizes for each data item, the formula used for calculating accuracy and the reports that would be produced. The basis for measuring and reporting accuracy was based on keying or coding a sample of values for a second time without the operator knowing that it was for QA. The basis of the system relied on consistency between at least two independent sources as a proxy for accuracy. For example, if automatic coding and a manual coder selected the same code, then that was counted as accurate. If two coders selected the same code, but it was different from the automatically coded selection, then the code selected by automatic coding would be counted as an error. If automatic coding, and two manual coders each selected a different code, then this would be counted as an error. as the number of records changed would be small.

Data capture samples were selected for each Census District (CD) – there are approximately 11,000 households in a CD. A similar process was followed for all types of data capture and for automatic and manual coding, although the sample sizes differed according to how often values for individual questions were expected within the CD.

The minimum sample size for high volume fields that needed coding (Industry, Occupation and Workplace) was 2%. There were 30 million responses for Occupation, 27 million for Industry and 22 million for Workplace. There are around 30,000 people in a

Census 2001 Review and Evaluation

CD and Service levels were set at this level of geography, for these questions, as they produced a sufficiently high sample to assess accuracy and error rates.

The remaining questions were low volume, with Country of Birth producing 3.8 million, Ethnic group 3.9 million, Religion 1.0 million and Address 1 year ago 4.7 million responses for coding. Responses to these questions were infrequent at CD level, and the sample rates required to produce statistically significant data would have been prohibitively expensive for processing. So the sample was 2% or 40 records, whichever was the larger and the sample was applied at Estimation Area (EA) level (around 500,000 people), and accuracy calculated and reported at that level.

The reports provided the confirmation that the sample sizes were in line with the DQMP, that accuracy was being maintained, or pinpointed CDs or EAs where accuracy had fallen below the SLA. These reports were the primary management tools for the LM and ONS data quality managers.

The ONS Checks on Data Quality

Data Capture

ONS carried out a thorough check of captured data following the 1999 Census Rehearsal. The check involved manually extracting data from paper forms or images, and comparing that with the data that was exported from the LM system. This was not just a check on the capture of the characters and marks, but of the checks, validation, and multi-ticking rules that formed part of the system requirements. This was a laborious task, but the results showed that accuracy for data capture was exceeding the service levels. The results provided confidence in the quality of the data capture process, and reflected the content of the Data Capture reports produced by the LM QA system.

Further rigorous testing of the system and the revised report layouts satisfied our requirements that LM were accurately measuring and reporting the counts of characters captured, the type of capture, and the accuracy achieved.

In addition to the QA carried out by LM, ONS developed a Data Quality Management System (DQMS). This system produced a variety of validations, counts and distributions for the tick box questions. The DQMS team could also interrogate on many other variables to identify anomalies in distribution, which could, in turn, identify systematic error in data capture.

The level of confidence gained from detailed checks of rehearsal data and the number of checks carried out by the DQMS was the basis for deciding not to carry out independent checks on the quality of data from the 2001 data capture process.

Coding

Although there was confidence that the QA system that LM had developed would reflect the counts and accuracy as measured, there was no method by which they could identify systematic error in individual codes. A systematic error in, say, Occupation could exclude an occupation title from the output data, skewing the data in a different code or distributing records over a number of codes. It was agreed that ONS take on responsibility for checking accuracy at the individual code level, and report errors to the LM coding development manager, who, in turn would examine the errors and agree changes to indexes and systems with ONS.

ONS built a system to extract a random sample of coded records for each question in each EA. The accuracy of the records coded was checked by a team at ONS who were specially trained in coding each of the questions on the Census form. The result of the independent check was recorded on the system and reports produced for each EA. If particular problems were identified, further samples and analyses were carried out to pinpoint the cause of the problem and agree corrective action.

The delay in the delivery schedule provided the opportunity to carry out a complete check (100%) on all questions for the first two EAs. These thorough checks identified the majority of problems and, in discussion with LM, changes were agreed to the system and indexes. However, the delay in the delivery schedule, also meant that some 33 EAs were being coded, or had completed the coding stage before corrections were implemented. The effect of this on the

Census 2001 Review and Evaluation

overall accuracy rate is shown in the tables in the full report, although improvements were made to the output accuracy by applying corrections to the 33 affected EAs. No significant new problems were found in subsequent EAs, and the system corrections, coupled with the increasing experience of the coders, helped improve accuracy over time.

Checks were carried out on each delivery of data throughout the processing timetable. The number of questions checked and the size of sample varied according to the delivery schedule, staff availability and the size of the EAs delivered.

Results

Data Capture

The table shows the results for the six service levels relating to quality of data captured from the forms. The quality of capture was excellent for marks and characters, and exceeded the SLAs but the accuracy of capture for form identity as a field was slightly lower than the SLA. This, and the other issues surrounding the capture processes are discussed in the full report.

Coding

All text responses were submitted to an automatic coding system, and if not successfully coded were referred to manual coders using a Computer Assisted Coding system. The following table shows the LM report on accuracy that covered the whole of the processing schedule, together with the ONS assessment of accuracy based on the results of checks carried out, after delivery, on samples of coded data. For all but two address questions (Enumeration and Workplace) the accuracy assessed by ONS confirmed that LM had met or exceeded the SLAs, confirming that consistency is a good measure of accuracy. The differences between the LM and ONS results are discussed in the full report, as are some particular problems reported for each question.

The volume of text responses for Country of Birth, Ethnic Group and Religion are relatively small as the majority of respondents ticked a box. There were no tick box response options for Industry and Occupation which accounts for the high volume of coding for these questions.

Type of Capture	SLA %	LM Reported Accuracy %	Type of Response
OMR	99.3	99.84	Tick Boxes
OCR Alpha	96.0	99.03	Alpha characters
OCR Alpha numeric	95.0	98.99	Address and postcodes characters
OCR Numeric	98.0	99.75	Numeric characters for hours, rooms, year last worked
Date of Birth	99.5	99.93	Separate SLA @ character level
Form Identity	99.995	99.68	Separate SLA @ character level

Census 2001 Review and Evaluation

Question	Number Coded	% of Population	SLA %	ONS Assessment Accuracy %	LM Reported Accuracy %	LM Automatic Coded %
Country of Birth	3,780,151	6.36	96.0	99.80	99.8	81.6
Ethnic Group	3,866,964	6.51	96.0	96.80	98.6	75.7
Religion	1,045,874	1.76	96.0	97.00	98.8	74.4
Industry	27,970,005	47.08	88.0	88.22	89.1	66.8
Occupation	29,611,276	49.85	88.0	89.04	91.1	72.2
Enumeration	1,651,445	2.78 (h/h)	100.0	98.51	99.3	88.4
Workplace	22,056,446	37.13	94.5	86.12	94.3	71.8
Address 1 year Ago	4,720,878	7.95	96.5	96.19	98.1	83.6

Conclusions

The quality of the captured and coded data was excellent. This is the first time that QA processes have been measurable and have continued as an automated background activity for the whole of the processing operation. Confidence in the system stemmed from a robust testing strategy developed within ONS, and carried out by ONS with a high level of commitment of system and staff resources from the contractor.

The QA system developed by LM and the ONS assurance processes worked well, complemented each other and delivered data that improved in accuracy during the processing schedule. All major problems in coded data were identified in a comprehensive check of the first EA. However, any future contract should incorporate a requirement to have a large block of data delivered early enough for checks to be carried out and corrections applied before much more of the data is coded.

Consistency is a good measure for accuracy except where descriptions are coded consistently, but inaccurately. Checking a sample of coded data was an effective method of identifying these systematic errors, assessing reported accuracy and improving accuracy over time.

The LM System Architect, Coding Development manager, Data Quality manager, and ONS processing staff provided continuity, experience and expertise throughout the development and operational stages. This ensured an ongoing commitment to delivering a successful QA system and to improving the quality of captured and coded data.

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