DATA COLLECTION INITIATIVES AND BUSINESS SURVEYS IN THE OFFICE FOR NATIONAL STATISTICS

Peter Thomas, Office for National Statistics, UK

Over the last seven years the United Kingdom Office for National Statistics has been implementing a series of initiatives to improve the collection and processing of business statistics data in the UK. The paper describes the recent history and covers proposals, which are at present either at a pilot stage or projected for the next four years. A range of new technology solutions have been applied to data collection. Document imaging and scanned forms have replaced paper forms for all processes. For many of the smaller inquiries Telephone Data Entry (TDE) is increasingly being used to collect the data. Pilots of data collection through the Internet have also been carried out. Having virtually all incoming data in electronic format has allowed the introduction of workflow systems across a wide range of data collection activities. The paper covers the future strategy which will primarily centre around the development of data collection on the Internet and changes to the scanning systems that will allow data to be captured by question providing the potential for much greater flexibility in business form design. The paper also covers some of the major processing developments that have focused on automatic and prioritised editing.

1. Background

1.1 The Office for National Statistics (ONS) conducts a range of business surveys which form the basis for economic statistics used by the Government in its monitoring of the United Kingdom economy. These surveys are a combination of short term - monthly and quarterly - surveys, often to identify turning points in the national economy, and annual surveys that generally measure levels. Examples of these surveys are:

- Monthly Wages and Salaries Survey used as the basis for the calculation of the Average Earnings Index
- Monthly Production Inquiry used as the basis for the Index of Production
- Monthly Inquiry into the Distribution and Services Sector used as the basis for the development of an Index of Services
- Quarterly Capital Expenditure Survey to measure levels of capital investment in the economy
- Annual Business Inquiry (employment) used as the basis for employment estimates
- Annual Business Inquiry (financial) used as the basis for a range of economic aggregates that feed through into the National Accounts
- New Earnings Survey used to assess earnings by occupation, gender and location.

1.2 There are about 100 business surveys each year, the majority of them conducted under the Statistics of Trade Act 1947. Currently 1.8 million forms are sent out each year to more than 320,000 businesses.

1.3 Most business survey samples are selected from the InterDepartmental Business Register (IDBR), a comprehensive register of businesses and their structures. The major sources of updating of the IDBR are from Value Added Tax and Pay as You Earn records that identify new and closed businesses, and from the Annual Register Inquiry carried out by ONS to gather local workplace details for multi-site businesses. Sample selections are usually taken from IDBR populations stratified by industrial activity and by size measured by employment. The chance of inclusion in a survey increases with the size of a business. The largest businesses are always included in relevant samples because of their importance to the economic statistics. To ensure that any business survey results are representative, businesses of all sizes and types of business activity have to be included in the survey samples.

2. Overview of the Business Survey Data Collection Innovation Programme

2.1 ONS has had a Data Collection Innovation (DCI) Programme in place since the mid-1990s in support of an overall data collection strategy for business surveys. This strategy has three main aims in relation to business surveys:

- To reduce the form filling load on businesses contributing to the surveys
- To improve the quality and timeliness of the business data provided
- To reduce the cost and increase the efficiency of business data collection

2.2 At present the main operational elements of the business survey data collection process are:

- Paper-based systems for all surveys, using image scanning and intelligent character recognition (ICR), which has very largely eliminated paper handling beyond the scanning phase
- "Touch-tone" telephone data entry (TDE) for a number of short term inquiries where the amount of data collected on each form is small
- Workflow applications in Lotus Notes to manage the data processing and validation tasks
- Central receipt and indexing of faxed returns which are then redirected to data validation, as opposed to data capture, staff for online data take on.
- A Contributor Comments Database (CCD) on which any relevant information gained from contributors when they are contacted, for example to query their returned data, is recorded

2.3 The data collection programme has made significant progress over the years on all three of the aims above. There has been a reduction in compliance as the CCD has lessened the need to contact contributors to query their data. Progress has been made on the more timely provision of data through the use of TDE which is now the collection method used by businesses for over 20% of business forms. TDE has also had an impact on quality through the ability to have some limited validation built into the system at the point of data collection.

2.4 The DCI Programme has helped to achieve significant improvements in survey processing efficiency. In this context the use of scanning and ICR, in combination with organisational changes, has led to reductions in excess of 50% in costs. These organisational changes have progressively moved from a structure where statistician led commands managed all aspects of a survey from capture to results, to one where while the results and data collection functions have been separated. Virtually all data collection for business surveys is now concentrated in a single Division, with other Divisions responsible for results. Finally two successful pilots of collection of business data over the Internet have been completed, and further work in this area will be taken forward as part of an overall Modernisation Programme for the ONS.

2.5 Over the next few years the main priorities of the next phase of the DCI Programme will be to reduce the burden on contributing businesses and to improve the quality of the data provided, including its timeliness. The Programme seeks to achieve this through the following vision:

- Maintain-paper based systems for each business inquiry using scanning and ICR for data capture. While the volume of paper based returns is expected to reduce, a significant number of businesses are expected to want to continue to provide their data through this medium

- Further take up of TDE in surveys where it is already in place and roll out to other appropriate business surveys
- Develop and implement Internet data collection systems
- Integrate data capture of faxed returns into the scanning and ICR processes.

2.6 The business benefits of this DCI vision can be summarised as:

- There will be timeliness benefits where businesses switch to the electronic options of TDE and the Internet. For example, in some of the monthly surveys up to 6% of forms are returned within the two days after the survey closure date
- There will be quality improvements through the ability to carry out data validation at the point of data capture for both TDE and the Internet.
- Data validation at source will also reduce business costs as there will be less need to contact contributors to query their returned data
- Web based data collection will provide a secure means for ONS to receive data from the increasing number of contributors who want to return their data through use of Email
- There will be reduced data collection costs through both a reduction in printing and postage costs, reduced data capture costs where it is returned electronically, and reduced validation costs with more validation carried out at source.

2.7 There is no reason why the business benefits outlined above could not be achieved with systems specifically developed for each of the three main data collection mediums - paper, TDE and the Internet. However, the current DCI Programme goes beyond this by moving from the current form/page based recognition, which is very inflexible and resource intensive, to question based recognition. The ultimate aim is to automate the process of business "form" production irrespective of whether the form is paper, telephone or web-based.

2.8 The move to question based recognition also has other significant benefits. It would support the development of customised or bespoke forms which would provide much greater flexibility in their design. This opens up the potential to markedly reduce compliance costs to businesses. It would provide a ready facility to incorporate one-off questions into business inquiries and also the ability to enable capture from faxed returns to be integrated with that for paper forms.

2.9 The following sections of this report provide more details on the evolution and plans for business data processing within the ONS. It does this in four sections.

Section 3	Data Capture from Paper Forms (scanning and ICR)
Section 4	Telephone Data Entry
Section 5	Internet Data Collection
Section 6	Integrating data collection across paper, TDE and the Internet

2.10 The above focuses on improvements to the business survey processes by which businessses are able to return there data to the ONS. Section 7 briefly refers two processes, automatic and selective editing, that have been developed to reduce the amount of editing of the returned data, but without impact on the quality of the survey results.

3. Data Capture from Paper Forms (scanning and ICR)

Document Imaging

3.1 The first document imaging project was set up in 1995 and tested on the Monthly Turnover Inquiry. The successful pilot allowed document imaging to be rolled out to the remainder of ONS business inquiries. This has now been applied to 95% of business forms received in the ONS.

3.2 Kodak Scanners with OCR_for_Forms software are used to take images of the business survey forms. These images are then used to enable the form-type to be recognised, through a form identity number, and then to capture the hand-written data. As part of this process data that cannot be read is flagged and this is then routed to data perfection staff, along with the image, using UNIBASE software. This system operated at a data character recognition rate of around 97%.

3.3 In 1998/99 "Drop-Out Colour" technology was implemented, whereby coloured data entry boxes drop out at the form scanning stage to allow the software to recognise only the new data added to the form by the contributor. This has increased the data character recognition rate to approaching 99%, with less resource required for manual verification. All forms, with the exception of a small number of long complex forms with a great deal of free format responses, are now processed by the system. The system has proved to be reliable and relatively easy to use. Following the introduction of drop-out colour a move to standardise on white paper for business inquiry forms has further improved the quality of scanned images and the data capture from returned forms.

3.4 The bulk of incoming forms returned by fax now arrive at a central fax server. A simple indexing system enables the electronic fax to be stored along with the images of returned paper forms. The appropriate inquiry processing section is then electronically sent the form, and currently these staff key the data from the fax image.

Use of Lotus Notes

3.5 The organisational changes in data collection and the move towards electronic versions of data, whether document images, TDE or Internet, gave rise to demands and opportunities for improvements to working procedures. Lotus Notes was selected as the medium for the introduction of new workflow systems. Apart from the widespread adoption generally for this kind of application, Notes is also used by the other national statistical institutes providing a ready made pool of relevant experience which the ONS could share. The initial pilot systems proved very popular with the users and rapid growth in the user community and the range of applications proceeded faster than our ability to manage the expansion. The strengths of the Notes applications were that they reinforced the move away from single inquiry workgroups and allowed processes, which were similar for inquiries, to be dealt with in a consistent way.

3.6 A major success has been the Contributor Comments database, which allows all inquiries to share the soft information arising from comments on forms, conversations with contributors or desk research into business organisation. This has reduced the effort required to explain unusual data and eliminated duplicate telephone calls from different inquiry areas to a single contributor.

3.7 Notes is also an excellent tool for work allocation, prioritisation and monitoring, which has eased the management of work areas with responsibility for a mix of short term and annual inquiries. Some of the work arising from a form, such as change of address or business structure, is more appropriate for the business register than the inquiry and the combination of Notes and electronic sources of data makes it easy to reassign work in a secure and audited environment.

3.8 The problems with the initial Notes applications were partly the problems of managing the rapid expansion, but a major technical drawback was the difficulty of communication between Notes and the legacy inquiry systems, which remained vital for the number crunching processes such as validation and analysis. Systems became hybrid, with users having to move between old and new systems and some bulk data being transferred to Notes where it was no longer being updated in real time.

Proposals for Document Imaging

3.9 Document imaging is now being developed in the second stage of the DCI Strategy. Over the next two years proposals will include the improvement of the Intelligent Character Recognition process, with a move to question-based recognition rather than the current form-based recognition. Data capture will move to being by individual question rather than for a set of questions fixed by the template for each page of an inquiry form, thus giving greater flexibility in the design of the questionnaire and reduced maintenance costs. The new methods will also allow the procedures used for handling paper forms to share common systems and standards with other media, via the Collection Database which is discussed later.

3.10 Improvements in the processing of faxed data are also planned. The new question-based recognition software currently under development will allow the integration of the output from the central fax server with the ICR process used for paper forms. This will enable automated data capture straight into the inquiry processing systems, and removes the need for online data take-on by the inquiry staff.

4. TDE "Touchtone" telephone data entry (TDE)

4.1 Data from forms with only small numbers of variables are increasingly entered via telephone data entry technology. In 1995 the first pilot of telephone data entry was carried out for the collection of data for a new Service Sector Price Index. Contributors returned price information on a range of products. New contributors were offered telephone data entry as the means for supplying data for the inquiry.

4.2 Telephone Data Entry uses the tones of a telephone keypad to make responses and to allow the contributor to undertake a dialogue with a set of recorded messages. Contributors may optionally add voice messages to explain validation or credibility problems.

4.3 Contributors received a routine hard copy contact letter each period requiring them to provide the price for a stated product or products. With TDE, this requires a contributor to enter a reference number and then a price for the product(s). There is also a facility for a contributor to leave a voice message which can be played back by the survey processing staff. The system allows for data entered to be automatically checked against a previous return, including a prompt for a comment from a contributor where new data entries are inconsistent with previous data, thus providing a vital validation check.

4.4. TDE was extended to the Producer Price Index, where the majority of price quotes per month are supplied this way. There is an option in the collection of data for the Retail Sales Inquiry, which enables contributors to choose whether to supply data on the paper form or using TDE. The facility to use TDE is now offered on over 40% of short-term inquiries. A recently launched Vacancies Inquiry has virtually 100% of its data capture via TDE.

4.5 Initially there were problems with capacity, in dealing with comments provided by contributors, and to a lesser extent within the telephone network. These were all been ironed out and the system was very reliable and continued to run successfully over the "Year2k" period without intervention. However, the popularity of TDE as a data entry mechanism created its own problems within the ONS because the original system had only a limited number of contact telephone numbers. The hardware of the system was therefore reviewed to increase the overall capacity of the system, and to provide additional facilities.

4.6 The original system ran on non-standard hardware using a non-standard operating system with analogue telephone lines. Since the original system was purchased the market and technology had moved on considerably. The system was therefore replaced, which met the following objectives:

- move to a Windows NT platform;
- expansion of system capacity;
- resilience of the system;
- increase in the capability of the system;
- provision of desktop capability, including inter-connectivity with ONS' telephone handsets (although this is not yet in use);
- provision and use of ISDN (digital) capability;
- reduction in the maintenance cost per line of the system, although it is recognised that an increase in capability may mean an increase in maintenance

Future Proposals for Telephone Data Entry

4.7 Now that the system has been upgraded the inquiries using the system are being expanded. Telephone data capture through TDE now accounts for over 20% of returned business survey forms. A roll-out programme is underway to offer TDE on other suitable business inquiries, with expectations that the percentage of forms returned by TDE will increase to 30% . This roll-out is now using dialogues that have been standardised. TDE is also used to provide other facilities, for example to order a duplicate form, and further similar applications are under consideration. The use of voice recognition will also be considered.

4.8 Investigation will be made into the business requirement for Computer Assisted Telephone Interviewing (CATI) and Computer Telephony Integration (CTI). For example, applications could include assisted dialling for response chasing or a purely telephone-based inquiry.

5. Internet Data Collection

5.1 Data collection via the Internet was piloted between 2000 and 2002 on two business inquiries to ascertain whether the concept of Internet data collection would be successful for business inquiries. This work was stopped in 2002 as it was realised that the technology to develop the next generation of Internet data collection systems within ONS business surveys would be different from that used for the pilots. Proposals for the future work on Internet data collection for business surveys are expected to be taken forward as part of ONS's Modernisation Programme. A key focus of this Programme is the standardisation and systemisation of processes used within ONS for business surveys, and also across social and administrative data capture where appropriate.

5.2 The purposes of the two pilots were similar in nature in that as well as proving the technical feasibility of Internet they would also provide an assessment of the potential and benefits, including efficiency savings, from the use of Internet for data collection. The lessons learned from these pilots would enable recommendations to be made that would guide the further development and implementation of Internet data collection. Finally they would enable ONS to make progress towards E-Government targets of all business surveys having electronic options by which businesses could return their data.

5.3 The development process for the pilots was similar starting with contact with contributors to establish how many would be prepared to take part in the pilot. This was followed by the development of prototype systems, which were subject to amendment arising initially from the testing process, and then from feedback from users of the system and contributors in the pilots. Guidelines for use by both the contributors and the staff that would process the returns were compiled and detailed training given to the processing staff. Prior to the pilots going live, agreement was reached on the information needed to support their evaluation, including whether there were any modal effects on the data. These were based on control groups of contributors who had similar characteristics to those in the pilots. The main difference between the two pilots was that the second one to be set up included a registration system as part of the arrangements to manage the security of the data.

5.4 The points below summarise the main lessons learned during the pilots:

- Not all contributors have internet facilities and requested paper forms;
- There were still issues of non-response and paper forms were required for enforcement purposes;
- The "feedback" database was successful as a way of gathering contributors views;
- The web was quick and user friendly;
- Problems identified by contributors were accessing the system using their user ID and password, browser problems and server availability;
- The Registration process created delays in response;
- The on-line validation in one of the pilots was thought excessive by the contributors;
- No print facility for contributors for previous and current returns;
- Problems when contributors closed down system and tried re-entering to amend data.

5.5 Information on the level of take-up of Internet as a data collection option by businesses comes from both of the pilots. In the first a decision was made to increase the number of Internet contributors. This proved to be a time consuming exercise with 130 contacts made with contributors in order to get a further 39 involved in the pilot.

5.6 The second pilot provides information on the extent to which businesses will use the Internet once they have expressed an interest in providing their data by this method. Of the 126 potential Internet contributors,102 (81%) provided some feedback over the Internet with 83 (66%) providing acceptable Internet returns. For the others there were problems with difficulty or inability to register and some not in fact having Internet access.

5.7 In terms of the staff processing the Internet returns the view was that after a period in which a contributor had to settle into the use of the Internet in providing their data they preferred processing data collected via the Internet. While it is recognised that the pilots were quite small in scale, investigation found no evidence of modal effects for data supplied over the Internet. Internet also provided a direct benefit to contributors in that a "reward" system was built in with contributors able to access screens that allowed them to compare information for their business with trends across the industry as a whole.

5.8 As part of ONSs modernisation Programme, future DCI work, including that on the internet of the collection tool is expected to:

- Recognise and incorporate the ONS corporate dimension and work as part of corporate projects for technical development of all aspects of the DCI programme including the Internet;.
- Incorporate appropriate registration and authentication;
- Replace returns of unstructured data by a "secure" E-mail system.
- Follow design standards for data collection instruments;
- Take on board "lessons learned" and recommendations from the pilots.
- Work with JAVA and Oracle the corporate tools, to collect data into the Corporate ONS Repository for Data (CORD), including integration with CORD irrespective of the means by which data is collected. The Business Collection Database referred to below is expected to be integrated within CORD.

5.9 As it currently stands, the overall DCI programme and of which Internet data collection is a major part, is expected to reduce business survey processing costs by £800K (some 10%).

6. Integrating data collection across paper, TDE and the Internet

6.1 In the future it is planned that data for business surveys will be returned by paper form, fax, TDE, or by Internet. Developing and implementing these as separate systems implies large overhead, maintenance and data integration problems. We expect to solve these problems with the introduction of the Business Collection Database.

6.2 The Collection Database will be used to drive all the processes connected with the capture of data for inquiries, integrating the various collection media. It will underpin the collection of data, irrespective of the medium by which that data is collected - it will cater for data returned to the office on paper, by fax, via the Telephone Data Entry (TDE) system, and via the Internet. Eventually the database will also drive the production of the collection instrument; ultimately it could potentially be used by the Data Validation Unit when validating and editing returned data.

6.3 Benefits of this approach include:

- a single integrated system supporting and driving data capture via all types of media, and managing increasing mixed mode data collection in a standard way;
- increased commonality and standardisation of processes, and therefore reduced maintenance costs;
- the provision of considerable data for the management of the collection processes, including for audit and monitoring processes;
- one of the essential foundations that must be in place before a "Bespoke Forms" production system can be implemented for the fast and flexible production of forms; which leads to increased flexibility for customers to change inquiry questions and the ability to vary questions within inquiries and a reduction in compliance costs;

7. Methodological work in support of business data capture

7.1 In spite of attempts to use best practice design questionnaire techniques to minimise the number of errors, respondents still make mistakes. These can be as simple as returning data in pounds instead of in thousands of pounds, or more complicated through misunderstanding the questionnaire or the inability to provide data as requested.

7.2 Validation techniques are built into most of the ONS survey systems to identify such errors and also to identify suspect data. Predetermined validation gates are used to identify either large changes in data items. This can be by comparing individual data items with those returned for previous periods, or for new respondents, comparing their data with those expected for a particular industrial classification or employment sizeband. The validation system will also identify missing data or errors, such as incorrect dates, totalling errors. Some surveys also employ congruency checks comparing similar data collected from administrative sources, such as held on the business register , and from other surveys, such as 12 months turnover from a short-term survey with that collected for an annual survey.

7.3 Validation consumes a substantial proportion of survey resources - generally thought to account for up to 40% of total survey costs. Traditionally, National Statistical Institutes have believed that focusing a large proportion of resource on data editing and validation produces high quality survey data. However, over the last decade the old philosophy has viewed as inefficient, leading to high survey costs, high respondent burden and possibly poor quality data due to errors being introduced into the data through over-editing. The new philosophy advocates that good data quality is not guaranteed through large amounts of data editing and that valuable resources should only concentrate on the "important" suspect values.

7.4 The ONS has been exploring the introduction of new editing techniques such as selective or priority editing. The project concentrated on short-term improvements to existing systems, rather than replacing them. The aim was to develop and apply two new methodological approaches to improve the efficiency of data validation and editing without adversely effecting data quality.

Automatic correction of systematic errors

7.5 Satisfactory methods of automatic correction were developed for two kinds of systematic errors:

- thousand pounds errors, in which respondents give information in pounds when asked to provide it in thousands of pounds;
- totalling errors, when there is inconsistency between a reported total and the sum of its reported components.

7.6 The automatic editing method for thousand pounds errors focuses on the main variable value returned at respondent level, in this case, total turnover. The method operates by comparing the returned value in the current reporting period with the accepted value from the previous reporting period. Where the ratio of these two values falls within a specified range centred on 1000, the returned value is adjusted by dividing by 1000 and rounding the value. The automatically edited value is then used in all further data processing.

7.7 The automatic totalling method was introduced for employee totals. It focuses on the total employee value, returned at the respondent level for business surveys and at the local unit level for the business register. The method compares the sum of the components for the current reporting period with the accepted total employment value from the previous period. If the difference lies within an acceptable range, then total employment for the current reporting period is amended to equal the sum of its components. The automatically edited value is then used in all further data processing.

7.8 Both methods of automatic editing operate before the data are passed through the survey specific validation system, hence reducing the errors that are triggered by the validation system. This in turn demands less editing resource and ensures that this type of error is edited less subjectively. These methods were successfully piloted on several live survey processing systems and have been introduced to all suitable surveys.

Selective editing

7.9 This prioritises validation failures so that only those for which correction is expected to have a material impact on the survey outputs are followed up for editing**.** This process takes place as part of batch data take-on and follows automatic editing. Data changes arising from many validation failures result in negligible changes on the survey estimates. Selective editing allows for individual businesses data that fails standard system validation checks to be assessed and scored according to its level of impact on the survey output.

7.10 The possible application of selective editing to business surveys in ONS was investigated and analysed, using four months of data from one key ONS business survey, by the University of Southampton (UoS). The UoS methodological evaluation study identified the key variables that required scoring according to its level of impact on the survey output. Individual businesses data with scores above a predetermined threshold were identified as requiring scrutiny and editing, and those below the threshold were accepted without further scrutiny. The pilot of selective editing on one ONS business survey resulted in a reduction in the number of data items requiring scrutiny of over a third. The effects on the survey results were confirmed as negligible.

7.11 Selective editing has now been rolled out to all appropriate short term surveys. Although the principle remained the same this did involve identifying specific key variables for the scoring of each survey's data. Studies are underway to investigate the possible application of selective editing to longer term and more complex surveys. This will take time as piloting has demonstrated that the same method cannot be simply transferred to another survey without impacting on the quality of the survey estimates.

7.12 The introduction of selective editing led to the development of graphical editing to support selective editing and to provide DVB supervisors with a tool for quality assuring the work of the data analyst. During the pilot of selective editing it was noted that errors in the non-key variables may slip through the selective editing process, as the score only takes account of the key variables. Some additional form of editing was therefore required to check for significant errors in the unedited values for the non-key variables. Graphical editing performs this role by the plotting of graphs by industry, so that the analyst builds up a knowledge of the industry, recognises patterns and relationships and therefore detects anomalies in the data. Supervisors can also focus their quality checks by graphically determining the large changes in the weighted and unweighted contributor data.

8. Conclusion

8.1 The ONS has made considerable steps over the last seven years in improving its business data collection processes. Further challenges remain ahead including the introduction of web based data collection and the ability to generate bespoke forms. Much of the change over the next three years will be through the ONS's Modernisation Programme.