Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

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

Peter Thomas and David Baird[1]

ABSTRACT

Over the last five years the United Kingdom Office for National Statistics has been implementing a series of initiatives to improve the process of collecting 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 some inquiries the form has been eliminated by adopting Telephone Data Entry (TDE). 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 also covers the future strategy on TDE data collection on the Internet, and covers current pilots and security issues under consideration.

1. INTRODUCTION

Since 1993 the UK Office for National Statistics (UK ONS) has been implementing a series of initiatives to improve the process of collecting data on business statistics in the UK. The objectives of this work are to reduce the costs of data collection, reduce the burden on data contributors and improve the quality and timeliness of data collected. This paper will describe the recent history and then go on to cover proposals which are at present at a pilot stage or projected for the next four years.

The stated objective of Data Collection in the UK is **"to get the collection right first time, with validation at source providing accurate data and appropriate commentary"**. These aspirations are very important to the way data collection is undertaken in the UK, and will be reflected in the paper.

2. RECENT HISTORY

Since 1994 the UK ONS has embarked on a programme of change in the area of data collection activity. The programme aimed to do many things. First of all there was a wish to brigade together all the activity in data collection and validation within subject areas to create a horizontal split between data collection and validation on the one hand and the production and publication of results on the other. The intention was to develop the structure where better use could be made of specific expertise in data collection and in the production of results. The second target was to deliver efficiency savings in the collection and analysis of data, in order to provide funds for new initiatives, particularly in economic statistics. The final aim or bi-product was to create a "paperless office" where all data were computerised from the moment of entry into the office and to remove the need for storing hard copy documents. All these objectives were to be achieved against a background of minimising the burden on contributors which form filling imposes.

[1] Peter Thomas, Head of Business Data Division, Office for National Statistics in the UK and David Baird, Head of Data Systems Branch, Office for National Statistics in the UK

The creation of the Data Collection Units (DCUs) as they were then called involved significant change in the way the Office operated leading to the following benefits.

- Savings of between 40 and 50 per cent were achieved in the whole process;
  - Staff were able to concentrate on the contributor relationship as well as producing results.
- The need for better data capture techniques became apparent;
  - It became the responsibility of some staff to capture and validate data only and to view the operation from that perspective and this led to better operating practices.

3. DATA COLLECTION INITIATIVES STRATEGY (DCI STRATEGY)

Shortly after the creation of the Data Collection Units, pilot projects were set up to test the feasibility of capturing data electronically. The following pilots were set up:

- **Document imaging**
- **Touch-tone telephone data entry**
- **Accountancy Software**
- **Use of Lotus Notes**

Over the 5-year period during which pilot systems were fully implemented, annual costs were reduced by £1m (for a total investment of £800k over a 4 year period). Each of the data collection initiatives had different levels of success, and is described in more detail below.

3.1 Document Imaging

The first document imaging project was set up in 1995 and tested on the Monthly Turnover Inquiry. The successful test 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.

The system is based on Kodak Scanners using OCR For Forms software which uses the scanned image to recognise the form and capture the hand written data. UNIBASE software allows data correction where amendment is necessary and these are linked by ROOT3 software. In addition, KoFax image controls enable staff at the desk to view the image on conventional seventeen inch screens alongside the databases which are generally stored in INGRES based systems on SEQUENT servers operating on a UNIX platform.

The images are transferred to a database accessed by Data Analysts who can toggle between form images and database systems. Data are also captured from the images and transferred directly into the transaction database which can be initially updated by verifiers (staff who amend those characters not recognised by the software). So far the system operates at a data character recognition rate of around 97%.

Recently more use is being made of 'Drop-Out Colour' technology, whereby coloured data entry areas 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 verification. All forms, with the exception of long complex forms with a great deal of free format responses are now included on the system. The system has proved to be reliable and relatively easy to use. A recent move to standardise on white paper for business inquiry forms will further improve the quality of scanned images and the data capture from returned forms.

Document imaging is now being developed for a second stage of the DCI Strategy. Over the next 2 years proposals will include

- A better integration of the development and management of forms design, scanning and Intelligent Character Recognition (ICR) processing and the workstation presentation of images. This will reduce the maintenance time and effort needed to make even minor changes

to forms. This is important as the centralised forms processing strives to become more flexible and responsive to customer demands.

- Improvement of the Intelligent Character Recognition process. This will include extension of the use of drop-out colour, the use of barcodes on every side of paper, i.e. every individual image, improved process controls, tracking of forms, images and data. 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.

- Consideration of the need for local scanning capability to enable "non-standard" contributor correspondence to be scanned and filed, for example contributor letters or company accounts.

- Improvements on the use of faxed data. The processing of incoming faxed forms must be investigated, with the ultimate aim of integrating the output from a central fax server with the ICR process and to enable data capture straight to the INGRES system. At present, incoming faxes to the central fax server are being electronically redirected to the appropriate inquiry processing section with data capture directly from the faxes not yet achieved.

3.2 Touch-tone telephone data entry (TDE)

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 applied in the collection of data for a new Service Sector Price Index. The contributors from the service sector produced prices on a range of products in industries such as office cleaning, bus and coach transport and private education provision amongst others. New contributors were offered telephone data entry as the main process for supplying data for the inquiry.

Telephone Data Entry (TDE), 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.

Contributors receive a routine hard copy contact letter each month requiring them to provide the price for a stated product. Usually this requires a contributor to enter a reference number and to identify the inquiry for which data is to be entered and then a price for a product. There is also a facility for a contributor to leave a voice message which can be played back to data analysts when necessary. The system allows for data entered to be automatically checked against a previous database and will also stimulate a comment from a contributor where new data entries are inconsistent with previous data, thus providing a vital validation check.

There are now over 2,000 contributors per quarter supplying data on this basis for the Corporate Services Price Index and the work has now been extended to the Producer Price Index, where the majority of 9,000 price quotes per month are supplied this way. In addition there is now an option in the collection of data for the Retail Sales Inquiry which collects monthly turnover from retail outlets and which enables them to choose whether to supply data using TDE or a hard copy return. The facility to use TDE is offered on over 40% of short term inquiries. A recently launched Vacancies Inquiry has virtually 100% of its data capture via TDE.

Although initially there were problems with capacity, with comments provided by contributors and to a lesser extent within the telephone network, these have now all been ironed out and the system is reliable. However the popularity of TDE as a data entry mechanism has created its own problems within the ONS because the existing system has only a limited number of contact telephone numbers. The hardware of the system has been reviewed to increase the overall capacity of the system.

The existing system runs on non-standard hardware using a non-standard operating system with analogue telephone lines. Since the system was purchased the market and technology has moved on considerably. The replacement of the system will meet the following objectives:

- to move to a Windows NT platform
- to expand the system capacity
- to ensure resilience of the system
- to increase the capability of the system
- to provide desktop capability, including inter-connectivity with ONS' telephone handsets;
- to provide and utilise ISDN (digital) capability;
- to reduce the maintenance cost per line of the system, although it is recognised that an increase in capability may mean an increase in maintenance.

With the system upgraded the inquiries using the system are being expanded. Some 20% of business inquiry forms have been replaced by telephone data capture through TDE. There are still a number of suitable business inquiries not using TDE. There are also potentially other areas of ONS, for example the Census Management Information System.

Investigation will be made into the business requirement for Computer Assisted Telephone Interviewing (CATI) and Computer Telephony Integration (CTI). Both of these will require digital capability from the desktop. For example, applications could include assisted dialling for response chasing and a purely telephone-based inquiry.

3.3 Accountancy Software

One of the first data collection strategy objectives prior to 1999 was to make use of existing accountancy software packages to collect ONS data. The vision was that key software developers in the accountancy field would be encouraged to include an option for contributors to output, automatically, returns required for national economic statistics.

Two firms were engaged in discussions. One firm took up the idea and explored the practicalities. The firm attempted to write a special ONS 'subroutine' into its package to allow automatic output of a number of statistics provided that a full accountancy database was held on the contributor's system. However the project faced a number of problems. The first was that each software supplier tends to have only a limited segment of the market and therefore only a very small proportion of contributors would access the option – thus it was not a cost effective commercial development for the software company.

A second problem was that 'Year2K' issues dominated work programmes between 1997-2000, and as a result resources to develop the statistics reporting elements were limited.

Finally, in the UK an agreed set of definitions and terms of all aspects of company accounts does not exist. As a result notes have to be very explicit and in the case of some companies derived variables are necessary to complete ONS returns.

ONS has recently re-visited the use of Accountancy Software and is working closely with Customs and Excise on a joint data requirement that will be put to software developers through the Business Applications Software Developers Association. The ONS philosophy is now trying to work with the data that businesses hold in their software packages rather than expecting exact matches to the detailed data requirements of ONS business inquiries. ONS is also looking at the role of ebXML and XBRL in accessing business data. There is no doubt that getting data directly from existing company finance systems remains a key objective which has the benefit of minimising the burden on contributors.

3.4 Use of Lotus Notes

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 Australian Bureau of Statistics, Statistics New Zealand and Statistics Canada, 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.

A major success was the Contributors' Comments database, which allowed 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 multiple inquiries to a single contributor.

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.

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.

.

Lotus Notes and workflow would feature highly in future developments, when these problems would be addressed

3.5 Summary of 1994-1999 period

The Data Capture Initiatives employed so far have achieved the majority of their objectives. They have removed a large proportion of paper that was circulated within the Office in 1994. They have generated huge efficiencies (over £1m on a total data collection budget of around £6m). They have led to the Office becoming efficient in other ways, particularly in the way that it is able to deal with contributors' conversations.

4. FUTURE

In 1999 the first phase of data collection was completed and the ONS moved to a second stage. At this stage all data collection and validation was brigaded in one division (Business Data Division) and a new 5 year plan for Data Collection Initiatives has been developed.

In the future the Office intends to develop a whole new range of Data Collection Initiatives. The history of Data Collection Initiatives in the UK ONS has been to provide inexpensive solutions based on costed business cases where substantial savings are required to justify investment in new technology. The success of the work so far has led to this second generation of data capture initiatives and the production of a new Data Collection Initiatives Strategy (available from Jessica Herbert@ons.gov.uk).

The main elements of the new DCI Strategy include:

- **Telephony developments**
- **Extending document imaging**
- **Internet based data collection**
- **Workflow and the use of Lotus Notes**

Again these elements have been costed and assumptions have been made about the likely efficiency savings which will stem from this work. The next 5-year plan, which will be completed in 2004/05, is expected to cost about £500k and generate total savings of around £500k per year from the third year totalling £1.8m over the full period. Developments in telephony and document imaging have already been covered in section 2. However, two other significant new developments have been added to the current list. This programme will be consistent with the ONS wide initiatives to improve its statistical infrastructure and information management.

4.1 Internet based data collection

Perhaps the most exciting new area of technology which can be applied to data collection in National Statistics Organisations is the use of the INTERNET to collect and validate data. In 2000 a pilot project was set up to allow contributors involved in the PRODCOM inquiry (25,000 contributors reporting on 4,600 products) to enter data over the Web. Initially the pilot is limited to 58 volunteer contributors and data has already been received using this method of electronic data transfer. This pilot has been successful, alongside a second pilot for research and development. Internet data collection is expected to be rolled out to other PRODCOM contributors and other inquiries once a comprehensive security and registration system is in place.

Early implementation of the Web based pilot will replace a paper based inquiry form with a Notes document which will be accessible from the Web. This will probably be the appropriate approach while a minority of inquiries and a small proportion of contributors are involved in Web collection. Once contributors supply most of their data by the Web it may be possible to consider combining inquiries into a single contributor system which feeds into multiple inquiry systems. As with other non-paper systems it may not be worth designing reminder and enforcement arrangements for the Web. This will be reviewed as part of the project. Contributors who fail to respond and require enforcement action will revert to the paper system.

In the PRODCOM pilot while the contributor is entering data, checks can be made on internal consistency for that form and comparisons with historic data from that contributor. The contributor can save invalid data and return at a later stage. The contributor will be able to provide an explanation either by picking from a standard list or providing a comment, which will clear the error status.

In order to provide the contributor with an incentive to use the Web and reduce the compliance burden it should be possible to provide some kind of reward to the contributor. This could be very simple. It could include providing links to other Web sites which might be of interest. More ambitious would be data tailored to what we know about the contributor. In the case of the PRODCOM pilot, the contributor is shown aggregate results from the inquiry for earlier periods comparing the contributor's data with SIC and whole inquiry aggregates.

Data collection via the Internet will use Lotus Domino software, making Internet collection an extension of the internal workflow system.

Depending on the success of the pilot project, it is envisaged that there will be a standard framework into which all inquiries will fit, rather than multiple inquiry-specific applications. Inquiry staff would use a Lotus Notes application to amend existing inquiry questionnaires, or to set up new ones. The entire process of creating the questionnaire will be driven by the user's entries, including for example:

- The definition of questions (with links to standard question libraries as appropriate);
- The definition of data items;
- The definition of validation checks;
- Any additional text required;
- Help text and guidance notes.

In the long term it is proposed that questionnaires should be contributor based rather than inquiry based. The aim is a single bespoke form for contributors; there may possibly be monthly, quarterly and annual forms, covering all inquiries for that contributor. This approach guarantees no duplication between inquiries, and helps with the problems of consistency, congruence and coherence. It also means a massive reduction in the contributor burden.

4.2 Workflow and the use of Lotus Notes

Workflow is defined as a system which ensures that the full range of manual and computer processes are optimised and performed in a more efficient manner. Incoming data are routed according to status, and work can be shared within a team.

It is proposed that Lotus Notes will become the main interface used by all areas within business statistics. It will act as a client, or front-end, for all appropriate applications. It is also at the heart of plans to collect data over the Internet. Specific issues which need to be addressed are:

- Use of workflow for dealing with input/recoding and validation of data
- Business Register/Inquiry database interfaces
- Improved Notes functionality
- Development of corporate standards
- Provision of audit trails
- Development of better management information system

A major technical stumbling block to increased use of Notes so far has been the absence of real time communication between Notes and the legacy systems based on the Ingres relational database. Within ONS there are decades of development in many aspects of inquiry processing such as selection, validation checking and reminding which ONS cannot afford to discard. Research into the problems is beginning to bear some fruit with techniques emerging which will permit bulk transfer of data where that is appropriate, real time exchange of data with an acceptable response time and the initiation of processes from Notes with a minimum of redevelopment in the legacy systems. Turning this technical research into practical solutions requires returning to the concept of pilot systems in selected areas.

One of the key areas for attention will be the Business Register. It is proposed that this will be made available to inquiry systems in a transparent and consistent manner. This will also provide a solution to the issue of access for other government bodies. In the UK, The Government Secure Intranet (GSI) has provided a secure environment for departments to share data using Internet type technology. In future it is anticipated that both GSI users and ONS inquiries will have access via Notes to the business register.

5. CONCLUSIONS

The experience of implementing new data collection techniques so far in the ONS has demonstrated that several aspects of policy in this area are key. These include:-

- In order to set up the projects it has been useful to set a 5-year deadline for planning and to make assumptions about technology for those 5 years.
- Strategies should be costed over the period and estimates of savings, however crude should be made.
- The use of pilot projects enables technology to be tested and gives contributors and users the opportunity to refine the projects.
- Good project management of pilots is essential. It is vital that users, technical experts and customers are correctly identified and contribute fully to pilots. This does mean that pilots can then be rolled out more easily.
- Pilots (e.g. telephony) can be inexpensive and there are often a range of system options at varying prices.

For further information the authors can be contacted on peter.thomas@ons.gov.uk and david.baird@ons.gov.uk