OUTPUTS FROM BLAISE SURVEYS

James Gray
Office of Population Censuses and Surveys, London, UK

1. Introduction

On the face of it, the method of data collection should not have too much effect on the output methods. If we consider the collection, processing and output stages of a survey then we might expect to be able to use any collection method we wanted — whether traditional paper, or Blaise, or some other method — with little effect on the processing and none on the outputs to the customer. Looked at in the abstract, output systems could be seen as modules that should work irrespective of the data collection method. Even if the data delivered by the collection module has a different format, it is still the same data; there is no reason for any changes to the way that results are output to the customer.

In practice there is a number of factors working against this. One factor is that customer expectations have changed. There is a general push towards faster and cheaper outputs. This applies particularly to computer assisted surveys. If we are getting 'clean data straight from the field', transmitted nightly over the phone lines, then why can the customers not have the data the next day? The vast speed improvement to the 'clean raw data' stage focuses attention on speeding up all other processes leading to output. Similarly there is a strong incentive to make the outputs more flexible to meet the customer's needs, and more accessible to a wide range of users.

Before considering the outputs to the customer, it is worth while looking at the outputs from the data collection stage that form inputs to later processing stages (variable derivation, imputation, grossing, analysis).

2. Outputs to processing systems: data structures

In IT methodologies, such as SSADM, systems are designed in a 'pure' way, looking at the inherent data structures. They are in theory ideal for central processing functions. In practice they do not always address the real needs and problems, especially when applied as a dogmatic formula.

On a typical Blaise survey, the systems are driven by the inputs. Certainly within OPCS, we have made the conscious decision that the Blaise instrument should drive the whole collection and processing system as far as possible. We are able to do this because of Blaise's strengths: easy authoring language, flexibility, excellent for input, and (especially) very good metadata provision. These features more than make up for the limitations that mean that for very large and complex surveys the data structures that come out of Blaise are not always in the 'ideal' Third Normal Form.

If the data structures coming out of Blaise are difficult to deal with, what can we do about it? There are three main routes available to us:

1. One option is to try to tune the Blaise instrument to give better structured data. It is possible to make some progress like this, but it is vital to remember that the prime purpose of the Blaise instrument is input. It is important not to compromise the data collection process. Nevertheless, a look at the instrument should be the first step.

2. Another option is to reformat the data into a 'proper' Third Normal Form. This will give very good and efficient data structures that are a joy to work with. The problem is reformatting the data and the metadata. As identified below, reformat programs are notoriously error prone. They also take a long time to write and form a barrier between different parts of the same system. For example, on a survey with a three month field period, one view would be that there are three months available from the start of interviewing to write and test the reformat program. After all the interviewing had been done, the data would then be passed through the reformat programs and made available to the derivation and analysis

programs. This means that for three months there is no chance to look at the data as it came back, or to test any of the final systems on live data. There is no opportunity to feed information back to the field force (eg incorrectly used codes). If any errors remain in the later processes it is likely that they will not be identified and corrected as soon as desirable.

3. A third option would be to accept the data in the shape in which it was collected. This may create complications in the later stages but does have some key advantages. As there is no reformatting process, there is no likelihood of reformat errors. The data are always in the correct shape for every stage once collected, so can be used in later systems immediately — this enables a move away from traditional large batch processing. Knowledge of the data shapes and structures is important — anyone who understands the data at any one point in the system will understand it in all others; this makes systems more resilient to staffing problems. As the data in all parts of the system is exactly the same shape as collected under Blaise, it is possible to carry metadata right the way through the system; it is possible to generate large parts of the later processing systems automatically, thus making data delivery faster and more reliable.

Another option which we have tried in the past when converting surveys from paper-based to CAI is to reformat the Blaise data to fit the existing databases. This would match the hypothetical ideal whereby one could treat input, processing and output as separate modules that could be interchanged at will. Here the Blaise plus the reformat procedure could replace the paper-based input and validation suite. In practice we have found that this does not work very well:

- someone needs to specify the reformat program, and someone has to write it. Experience shows that reformat programs are particularly error-prone. However rigorous the testing, it is very difficult to spot specification errors.
- the old paper-based data structures and variable definitions will have their own little quirks reflecting the particular peculiarities of paper-

based data collection. It is easy not to notice these if you have lived with them for a long time.

- you end up constraining the input systems — there may now be possibilities for better ways of collecting, processing and analysing the data, but if the processing and analysis is to be done the old-fashioned way, there may be little point.

A simplified example illustrates the point. We want to ask people in paid employment what their occupation is, and use it to work out social class and socio-economic group. If the respondent is not working, then we will ask what their previous job was (if they have had one recently), and calculate social class etc from that. The best way to implement this on a paper questionnaire (where there are constraints on the routing complexity) is by asking two sets of questions:

```
WORKING                          Are you in a job?
If WORKING=NO goto RECENT
THISJOB                          Details about current job
goto HEALTH
RECENT                           have you had a job recently?
if RECENT=NO goto HEALTH
LASTJOB                          Details about previous job
HEALTH                           set of questions on health.
```

This will give us filter variables WORKING and RECENT, plus also two large blocks of questions — THISJOB and LASTJOB. These are essentially very similar (they may differ in some details). From this information, we will derive the variable SEG (Socio Economic Group):

```
IF WORKING=YES
   SEG is derived from THISJOB questions
ELSEIF RECENT=YES
   SEG is derived from LASTJOB questions
ELSE
   SEG cannot be defined
ENDIF
```

The derivation for SEG is quite complex and has to be repeated for the two sets of questions. Parameterisation of the derivation might help, but this is seldom done in practice.

If we then wanted to convert the survey to Blaise, we would have the option to keep the set of questions essentially unchanged, or to have just the one set of job questions, relying on the filters to tell us to which job they referred:

```
WORKING                         Are you in a job?
If WORKING=NO
  RECENT                        Have you had a job recently?
ENDIF
If WORKING=YES or RECENT=YES
  JOB                           Details about job
ENDIF
HEALTH                          questions about health
```

The block JOB will refer either to the current job, or to the previous one, depending on the values of WORKING and RECENT. Detail differences can be dealt with by routing details (which would be too fiddly for a paper survey). The actual question wording will be slightly different ('were you working as an employee' instead of 'are you working as an employee') — again this is no problem in Blaise though it would not work well on paper. Less data space is used, there is only one set of variables and one set of checks. When deriving SEG, only one derivation needs to be specified, from the single block JOB.

If we were relying on reformatting the Blaise output to match the data structures we had been using on the paper based systems, we would have gained some of the advantages (simpler Blaise instrument for example), but we would still have a derivation for the SEG variable that depended on a large number of potential input variables depending on the informant's status. This is messy and potentially error-prone. It would probably be worthwhile, but would mean that the reformat program has more work to do (and more chance of being wrong).

If the data structures are coming out as too complex for easy analysis, then they can be restructured as appropriate. The mistake would be to use a paper-based input or processing format in the belief that it is necessarily the correct structure for a universal processing module.

3. Outputs to customers: data

What kind of outputs do we send to the customer? This will depend on the survey organisation and the type of customer. OPCS's Social Survey Division mainly runs surveys for other Government departments. As such, we need to produce a lot of straightforward data files. If the customer wants the data quickly, we may provide preliminary data 'straight from the field' before any office editing or variable derivation. It is not good enough to provide just the data — the customer needs to know what is on the datafile and where. Here the strengths of the Blaise metadata system are very important. The setup generator is an invaluable tool for producing human-readable documentation, as well as record schemas for a variety of packages.

With the setup generator, you only get out what went into the original questionnaire — it is very important to keep the outputs in mind right through the survey design stage. By default, for example, the variable label will be the first 40 characters of the question text. If you have a large number of questions that start off 'In the last four weeks, that is the four weeks ending Sunday $REFDATE did you...' then you will end up with a lot of identical (and fairly meaningless) variable labels. Blaise does offer the facility to override this default — this should be used if you want meaningful metadata at the outputs stage.

Another problem with straight data outputs can be the data structures. Here we are regarding the customer as we would regard the next stages in an internal processing system. When writing a Blaise questionnaire it can sometimes seem fairly arbitrary where to put the SUBFILE statement in relation to loops and tables. The difference is noticed readily enough at the output stage when you discover you have an unwieldy data structure. Again, thought about outputs is needed from the start.

Sometimes you end up with a data structure from the collection stage which is far from ideal. Maybe you erred in the questionnaire design, or maybe there was no real choice. If you have done a lot of processing on the data — derived variables, tabulations, etc — then you may already have addressed this problem. If not, then you may be leaving your customer to do so.

4. Outputs to customers: tables

For many organisations, the main form of data output is as fully analysed tables. For the customer, the important thing is to be able to answer questions about the real world. Sometimes the easiest way is to reach for a published report and leaf through it until the right table is found.

Sometimes it would be easier if the customer could have a wider choice of tabular outputs than is normally available in a paper publication, and could read the information straight into a spreadsheet or word processor. One solution would be to provide the whole dataset and let the customer tabulate the data as required. Often this is the right answer, but not all users of the data want to take on several megabytes of data and wait several minutes while it churns through the tabulator (and not all suppliers want to distribute sensitive information at the individual level).

The solution is to use an electronic publication package. STATview is a very good example. This CBS product looks like a tabulation package, but takes pre-tabulated data as input. This makes it very fast and very secure (as individual level data are not supplied). As less data are needed than with a true tabulator, more information can be provided on a given size of disk.

5. Outputs to customers: media

Even in terms of the physical medium for dataset delivery, the world is changing. A few years ago it was considered sufficient to offer data on IBM mainframe tapes. Everyone wanted to look at the data on mainframes, and all mainframes could read IBM tapes (or have them converted via a bureau) — it was a very good standard from that view point. Nowadays they are considered irrelevant (much processing is done on microcomputers that cannot read mainframe tapes), and unwieldy (although physically large, they do not hold much data).

If the new platform for survey processing and analysis is the microcomputer, then the standard microcomputer storage devices look attractive for data delivery. Perhaps the most widespread standard is the 3½" 1.44MB diskette. It is often dismissed as being 'no longer at the leading edge of storage technology', but its very maturity is its strength. Practically every machine has the hardware to read 3½" diskettes, the standard is adhered to very well indeed, and the medium is very cheap. Is it big enough? For abstracted data in tables (eg with STATview) it almost certainly is. Even when supplying complete data files, for most small to medium sized surveys, capacity is probably sufficient. Use of data compression software can dramatically increase the effective storage capacity. The rate of compression will obviously vary a lot depending on how efficient the original data files were for storage, but our experience suggests that you may get 1500 hours of interviewing into 1MB compressed data.

If floppy disks are not large enough there is a wide variety of alternatives — too wide in some respects. Quarter Inch Cartridge (QIC), 8mm Video (eg Exabyte) and Digital Audio Tape (DAT) all have higher capacities (250MB, 1.7GB and 4GB respectively). They also have varying degrees of presence in customer offices, and a wide variety of data encoding standards. These are all tape-based media, and thus suffer from lack of random-access and problems with reliability and resilience.

One view of the data medium of the future would be CDROM. This has some of the advantages of diskette in terms of random access of data. Most (though not all) organisations now have CDROM readers: they are anyhow quite cheap at around £300 each. The basic data encoding standard (ISO 9660) is very strongly adhered to, though the more specialised standards for multimedia applications have varying degrees of acceptance. The capacity, at around 600MB should be enough for most applications, though there will always be the exceptional dataset needing more than this. Until recently, CDROM has only been an option for very widespread dissemination to a large number of customers. It was necessary to go to an external agency to have a 'master disk' produced (anything from £1000 to £3000), then a production run of at least 100 (£3.50 per disk, reducing to £2.00 per disk for runs of 5000 disks). It has more recently become possible to purchase equipment for less than £5000 that enables one-off production of CDROMs for about £15.00 each. This makes the medium much more attractive for delivery and archival of data than before.

Where fast access to up-to-date tables is important, on-line access becomes a possibility. In an ideal world there would be good high-speed network connections with customers, and they could access information as if it were on their own computers. There are three main problems that delay the implementation of on-line systems:

1. Availability of high-speed network connections. The availability of high-speed public networks varies greatly in different countries. Where they are not available, then there will be constraints on the data that can be accessed on-line.

2. Availability of adequate software for on-line enquiry systems. The need is to make available to the customer not just the information, but also the meta-information that enables them to interpret it, and also the ability to locate the information of interest; if an organisation provides a large amount of statistical data it can be very difficult to locate relevant items.

3. Security. It is essential to ensure that the enthusiasm to provide easy and flexible access to abstracts does not lead to compromising the security

of confidential data. It is important not just to ensure that confidential data are secure, but also to be able to convince all interested parties that the data are secure. In many cases the only convincing security measure is to ensure that all external access is to a machine that has no physical or network link to any machines containing sensitive information. This could be a stand-alone PC (or group of PCs), or possibly to a computer in a different organisation.

The ideal solution would be a system that could run on stand-alone PCs, had good subject-searching facilities, and was capable of running in 'client-server' mode whereby a lot of the user interface could be handled by the client PC (this would reduce the load on the communications link).

One of the oldest methods of delivery of abstracted data is in the form of tables printed on paper. This is still very important, and will remain so for a long time. Paper outputs can reach a very wide audience, and are very flexible. Again, expectations have changed. The standards of layout and production that are expected are very high. Producing printed outputs to the required standards is demanding and time consuming.

6. Who is the customer?

The strong pressure from our external customers for faster outputs means that we need to be able to streamline the whole process. Our own outputs need to be produced with little human intervention by automating as many processes as possible. In addition, we need better control over what is going on in the field in order to make sure that we get good data back fast. How do we achieve better control? By better monitoring. How do we achieve this? By better outputs from the field systems for internal use.

This recognition that internal monitoring and management systems are in fact output systems in as full a sense of the word as the traditional outputs for customers, is a key step. In a conventional system, field management information (although seen as important by the field managers) can become marginalised as small add-on suites of reporting programs of fairly rigid structure. Once we recognise that we are looking at vital survey data, and that we have available to us all the tools that we would use for mainstream analysis and data provision, then we can make major steps forward. Our field managers then become not just service providers, but also customers to the survey.

One example of the kind of rethinking that can occur is in the use of data publication packages. In OPCS we have some very good Case Management Systems. These are built around a largely standardised case management subfile on our Blaise questionnaires and include a number of detailed reports (again largely standardised) that give progress and response information for individual surveys. The field manager for a survey can select an interviewing period to look at, and call up details for the whole survey, or separate field regions, or individual quota of work to get a very close look at what is happening. The reports are very good, but sometimes it can be difficult to get an overview at the organisational level. We have discovered that we can provide complementary information by making the case management data available in STATview. One of the great strengths of STATview is that it can take data from a large number of different input files and display them together. We can, for example, summarise the information for a particular survey month into a monthly file and then build all the recent months data into one publication. We can provide the same outputs for different surveys. The result is that we can provide a flexible system whereby field and project managers can choose to look at survey response rates over time, or across different surveys, or just for their own survey and month. The use of STATVIEW will not make the existing detailed reports redundant, but it does open up the ability to make better use of the information we are getting. Internal customers are empowered to create their own outputs without programmers, tailoring them to their immediate needs. They can also output the tabulated information in a way that will load directly into a spreadsheet in order to graph the data.

The internal customer has tremendous advantages over the external one. For example if the survey organisation has a PC network, then on-line access is no problem — the network will be very much faster than dial-up links, and there will be much less of a security problem.

7. Conclusion

The advent of computer assisted interviewing, and the Blaise software in particular, means that the data collection part of surveys is now fast and efficient. Attention should be turning to making this data available to later stages of processing and to the customer. There are many excellent tools that we can be using to help in this: some of these tools can be used to very good effect to serve the internal data customer.