

*18th Population Census Conference: 26 – 29 August 1998
Program on Population, East-West Center, Honolulu, Hawaii USA*

***THE USE OF
OPTICAL MARK READING (OMR)
FOR CENSUS DATA COLLECTION***

Presented by: Mr Kevin Orchard
DRS Data & Research Services plc
Sunrise Parkway
Linford Wood
Milton Keynes
MK14 6LR
United Kingdom

Telephone: +44 1908 666088
Fax: +44 1908 607668
E-Mail: census@drs.co.uk
Website: www.drs.co.uk

1. Introduction

Thank you for giving me the chance to address this conference. I should like to start with two admissions and one promise. Firstly, I admit that I am neither independent nor unbiased. I am the representative of a British company, DRS Data & Research Services Plc, that designs and sells Optical Mark Readers, and you represent a significant market for my Company's products! My second admission is that I am not going to talk about 95% of problems you have to tackle in running a census - I am going to concentrate on one very small but important problem, which my Company has a lot of experience of solving. Those are my two admissions; my one promise is that my talk will be short, to the point and as free from overt selling as my conscience will permit.

Most of you are involved in the entire range of activities that go to make up a modern day census. DRS have the luxury of specialising in a very small part of that process. It is however a part which seems to have caused an inordinate amount of trouble in some countries. It is also a key part in efforts towards yielding better value for producers and users of census data by providing faster results. The part of the process I should like to concentrate on is the stage between the data being collected and written down for the first time and the data being stored in a computer.

Every census goes through a stage where the data is first written on to a piece of paper. Equally, every census requires that the data that has been collected is entered into a computer. The reliable and validated transfer of completed form data to a computer system for further processing is the problem I should like to address.

There are only three practical methods of getting data from a piece of paper into a computer. These three methods are to KEY the data, to use OPTICAL MARK READING (OMR) or to use IMAGING combined in some way with OPTICAL CHARACTER RECOGNITION (OCR).

Before we compare these three techniques, I should first like to review each of these techniques and select the best way of using each technique for census data collection.

2. Alternatives

(a) KEYING

There are two components to any keying system. The first component is human. This is the component that does most of the work. The second component is the keying hardware. In the early days, data was keyed on to paper tape or punched cards, with clumsy mechanical devices. In those days it was the mechanics of the keying device that determined the performance of a keying system. With the introduction of key to disk systems the constraints on keying performance shifted from the hardware to the operator. As a result the rate at which keying is done, and the accuracy of keying is no longer significantly improving, although some minor improvement is possible with improved keystation ergonomics, computer aided keying and with the use of key from image systems. The latter can only realistically be used as part of an overall imaging approach but is the first new technique in keying for many years. It brings two main benefits, albeit at a price:

- Eliminating paper handling – there is no need to move, count, control or even store large quantities of forms.
- More effective and efficient key entry – as individual key operators can concentrate on one or a limited type of census question and double key entry administration can be done electronically.

However, to the best of my knowledge, no major technical developments are expected in human keypunch operators!

The main alternatives facing you in selecting an appropriate method of keying census data is the degree to which keying is centralised. In the past, because of the high cost of computers compared to peripherals, it had only been economic to centralise keying with a single powerful computer and tens or hundreds of key stations attached to it. However, with mass produced PCs and local and wide area networks the balance has long ago shifted in favour of each key operator having their own local computer. This in turn has removed the necessity for physically clustering all the key operators around a central computer. The possibility of

distributed keying has emerged. The general consensus appears to be that distributing the keying and moving the key operators closer to the source of the data is a good thing.

(b) OPTICAL MARK READING

Optical Mark Reading (OMR) is recognised as an efficient and cost effective data entry method for the large scale processing of category data. From its origins in examinations, it is now widely used in other large-scale exercises. OMR can relieve much of the effort and difficulty in keyboard entry and it has already been used in a number of census projects

An Optical Mark Reader (OMR) reads marks made on purpose designed forms and transfers the data straight to a computer system. It is a technology that is reliable, simple to operate, complementary to other data capture techniques and easily integrated with other elements of a computer based solution.

It should be recognised that an Optical Mark Reader can *only* read marks and barcodes and not numbers or letters or shapes. This is both a major advantage and also disadvantage of Optical Mark Reading. The disadvantage is clear but the advantage is that responses are simply recorded and immediately categorised. It is here that the use of OMR can aid the census process of large-scale categorisation of data. Indeed, the associated revision of the forms used for data collection into a format suitable for OMR can also bring its own benefits.

In use an OMR reader simply connects to an existing PC based computer system. A typical system configuration would be a number of OMR reading stations networked to a file server, with the associated printers, backup and power security.

(c) IMAGING AND OPTICAL CHARACTER RECOGNITION

The third method of getting data from a piece of paper into a computer is by imaging and then using Optical Character Recognition (OCR). The term ICR for Image or Intelligent Character Recognition is also in use. OCR has been “going to work” for the past 30 years and increasingly does work. However it remains a complex and expensive field and is the

most costly of our three techniques, keying, OMR and OCR. It is for that reason that I would like to look in more detail at the first two of these – keying and OMR.

3. The use of OMR for data entry

OMR can cost-effectively reduce the time required and increase accuracy in :

- Category data entry from household and individual census forms.
- The control, administration and logistics recording process.

The scale of improvement can be dramatic.

For example, as mentioned before, much enumeration data is simple category choice or numeric and such data is highly suitable for OMR. Other data may be enumerator or office coded to OMR. For overall process speed the number of office coding and write-in fields should, of course, be minimised.

The following table summarises data type and suggested completion suitability :

Data Type	Completion Suitability
simple choice and numeric	householder or enumerator
complex choice and simple coding	enumerator
complex coding and write ins (free text)	processing office

In many cases the overall data requirement can be achieved with a single double-side full page (12 x 9in, A4, or equivalent). Each page would be suitable for up to 6 or more household members and could also be used as a continuation sheet for larger households.

The advantages of a single page form are in reduced costs throughout the process - printing, distribution, completion, collection, office coding, reading and report production.

More detailed data requirements can be achieved with a two or more page booklet. It can be useful to divide a multipage booklet into pages of different data types :

- Primary data, to be completed by the householder or enumerator.
- Secondary data, consisting of office coded and write-in areas.

This has the advantage that primary data pages require no further office processing and may be separated and read initially to rapidly give a primary census abstract. Secondary data pages may be read later after the office coding processing.

Unique check digit form identifiers, such as barcode, may be used on census forms to aid the distribution and logistics processes and to match records when multipage forms are separated.

In addition to the main enumeration forms, OMR forms may be used as :

- Accounting forms - for batch control purposes. Such forms would be completed by enumerators on a census district basis and can be used to record location codes and other administrative details.
- Despatch/receipt forms - for recording the destination and receipt information for batches of enumeration forms sent into the field.

OMR forms are scanned accurately at high speed, the data is validated and available immediately together with error information and process statistics.

The use of keyboards and OMR is not mutually exclusive, in many situations a combination may provide the optimum solution. For example OMR can capture and make rapidly available 90% of the data from a form, leaving the keyed data to follow later.

4. The advantages of OMR in census data capture

Speed High-speed automatic readers, such as the DRS CD800 series and others, read over 8000 double-sided barcoded forms per hour. With real forms and less than ideal situations each scanner will still maintain a real read rate at 5000 plus sheets /hour, for hour after hour. With typical forms this is equivalent to over ½ million characters /hour.

Accuracy The combination of sophisticated readheads which can allow the reading of pen and pencil, accurate discrimination between intended marks and erasures, and the validation capabilities available in the system allows absolute accuracy and reliability. Rather than attempt an incorrect read, readers will reject forms in which there is doubt over the response.

Analysis Rapid, accurate and secure entry of data permits powerful analysis. This can be used not only for the generation of population statistics but also for performance reporting and the identification of discrepancies.

Economy Validated, accurate and secure data is quickly entered and at lower overall cost than manual or key only systems. Less time is required in administration and skilled resources can be released for more productive tasks such as analysis and report design. The economics of using OMR is further discussed later.

Capacity Valuable skills in data handling and analysis are built.

5. Comparison: Keying v OMR

Keying

Since you know roughly how many people the data will be collected from, and you know exactly how much data will be collected from each person, it should be fairly easy to calculate how many key stations, used for how long, are required to enter this data. The crucial figure you require for this calculation is your average keying rate (usually expressed in thousands of key depressions per hour). The figures quoted for this vary from as high as 15,000 key depressions an hour, down to 1,000 key depressions per hour. The high figure usually represents the rate at which a very good operator can key well prepared documents on good keying equipment after considerable practice (to put this into perspective, 15,000 key depressions per hour means pressing four keys every second, hour after hour!). The low figure comes from the more realistic route of taking

the total number of useful key depressions and dividing it by the total number of hours used to key them.

Most commercial data that is keyed is also verified (that is keyed twice). Whether this is necessary for census data is a decision that each country must make. Providing that the key operators are merely keying clearly written data, the accuracy of keying may not be a major issue. However, many census operations involve the transcription of the original data into a form suitable for keying. This transcription process can be significantly inaccurate and, of course, its inaccuracies are added to the inaccuracies of the subsequent keying.

The cost of keying is directly related to the speed of keying. We can consider the cost of keying in two parts. Firstly the cost of the hardware on which the keying will be done and secondly the cost of the people who will actually do the keying. The relative costs of these two parts will vary dramatically from country to country. For example, in Britain you can buy a keying station for about \$1,200, and you pay about \$10 an hour for keypunch operators. In some developing countries you might pay much more for the hardware and one-tenth the cost per hour for the key operators.

In order to arrive at an approximate cost for keying census data, we will need to make a series of fairly arbitrary assumptions. These assumptions are about right for a country with low labour costs and locally made keying hardware. We will assume that the keying equipment, usually a PC, can be purchased for \$1,000 and at the end of its census use can be sold for 30% of this. We will assume that the equipment is used for 12 months, with two 8-hour shifts, 6 days a week. We will also assume that allowing for training periods, sick leave, equipment malfunction, power cuts, etc., an effective keying throughput rate of 5,000 key depressions per hour is achieved. If we assume an operator cost of \$80 per month, we end up with a cost of about \$105 per million key depressions if not verified. If 100% verified, this figure would be \$210 per million key depressions. We will come back to these figures later.

There is no doubt whatsoever that keying and key operators are exceedingly flexible when compared with OMR or OCR. Human key operators are by far the most sophisticated data entry machines ever devised! They are able to compensate for inadequacies in form design and in form completion. However, this very flexibility can also be a

problem. In order to use this flexibility, key operators need to be given a level of freedom in what they key, which may in itself have serious drawbacks.

Optical Mark Reading

If we now consider the use of OMR we find that OMR readers make three types of errors. One is that they simply will not read a particular sheet. This, of course, does not produce inaccurate data – it produces no data at all. Assuming a mechanism exists for the operator to enter that data via a keyboard, this type of error only reduces throughput, it does not reduce accuracy.

The second type of error is that the OMR reader sees a mark that was not meant to be there. The ‘mark’ may be a fault in the paper, dirt or a poorly rubbed out mark. All OMR readers suffer from this problem. If you have a fault in the paper which is bigger, blacker and sharper than your intended marks, then the OMR read is going to read it. Whether this produces inaccurate data depends on how you have designed your form and how you validate the data. The frequency with which this happens is determined by the design of the OMR reader or more specifically with the design of its read head. Some crude OMR readers are very prone to this type of misreading. Other OMR readers are very good at discriminating between faults in the paper, rubouts and real marks.

The third type of error that OMR machines suffer from is failing to read intended marks. Once again, all OMR machines suffer from this problem occasionally. Again, the extent to which they suffer from it is determined by the design of the read head. Whether or not the error produces inaccurate data is determined by the design of the form and the validation applied.

You will notice that I have not included amongst the faults of OMR readers the transposition of one response into another. This simply does not occur with a correctly designed form and a good OMR reader. Substitution (as this type of error is known), is a feature of OCR readers and of key operators, not of OMR readers.

Now let us look at one of the most important criteria – cost. Once again, it is necessary to make a lot of assumptions in order to produce an approximate cost for collecting data via OMR. As far as possible I will make the same assumptions as were made in the keying example earlier.

We will assume that each OMR reader costs about \$30,000 and that a PC is purchased to go with it at \$1,000. Let us also assume at the end of the census this hardware can be sold for about 30% of its original cost. We will also assume that the equipment is used for 12 months, for two 8-hour shifts six days a week and that the forms throughput is a realistic 5,000 forms per hour. Although the operators of this equipment will not need to be key operators, we will nevertheless assume the same monthly operator cost of \$80. We will assume two operators per OMR reader per shift. For the OMR cost calculation we need one additional piece of information, and that is the number of characters that we get from each form. The figure I shall use here is 150. This represents the amount of data from a typical household on this type of form. These assumptions result in a cost of about \$7 per million characters compared to \$105 or \$210 for verified keying entry. In reality, one OMR reader under the above conditions would only take 4 days to capture the same amount of data as 2 key operators could process in a year.

Overwhelming evidence, perhaps. Other costs, of course, need to be considered and OMR forms do cost more and need to be handled with more care than ordinary forms. The overall cost of the exercise taking into account not just equipment and people, but also transport, storage, space and supervision is also important.

Other factors, such as speed, are also relevant. For instance, one of the few major census to take place in the last five years was in Australia, carried out by the Australian Bureau of Statistics. OMR and keying were used for data entry. Over 10 million multipage booklets were processed and the majority of the data released less than 12 months after the census date, and in fact 4 weeks ahead of schedule.

6. The future

Imaging systems continue to improve, particularly in those crucial areas of price and performance. One interesting development from the OMR field is in combined imaging and OMR. Such equipment would immediately capture and validate category data on forms, while saving write in areas as images for future processing. At this stage the form can simply be destroyed (or stored if confidence is lacking !).

A number of processes can then take place on the stored image. It should be noted that as each image will typically only contain a word or two or

some digits the actual size of each image in electronic terms is very small and therefore very manageable. For instance over a thousand such clipped images would fit on a single floppy disk and many millions on the hard disk of even the humblest office PC.

The stored image may be :

- Processed by OCR software to attempt to automatically recognise and convert characters or numbers.
- Used in a key or code from image system. The potential benefits of such a system were mentioned earlier.

7. Conclusions

In summary, I would like to suggest that OMR is a fast, accurate and cheap method for collecting census data.

Thank you for your attention.