PAPERLESS FAX IMAGE REPORTING SYSTEM(PFIRS)

Errol G. Rowe and Martin V. Appel
Errol G. Rowe
Statistical Research Division
United States Bureau of the Census
Washington, D.C. 20233-0001

INTRODUCTION: The Census Bureau conducts many surveys each month. These surveys involve anywhere from a few hundred to many thousands of respondents. Typically, questionnaires are mailed to respondents who are usually given a specified number of work days to fill them out and return them to the collection agency. The forms are then sent back to the Bureau where they are data keyed into a computer.

There are several inefficiencies in the current data capture process. First, consider the clerical workers. Theirs is a very tedious job requiring continuous hours of strenuous concentration. In order to ensure accuracy, a sample of the questionnaire forms is actually keyed by two clerical workers. Their entries are compared and those that match are passed. Those that fail to match are reviewed a third time.

The second problem with the traditional process is the delay caused by the mail. A recent Bureau of Labor Statistics study showed that up to 6% of questionnaires are received after the cutoff date for the initial estimates (Rosen, 92).

Finally, we consider the respondents. It is fair to conclude that most respondents can be placed into one of two categories; i.e., good or bad. Good respondents immediately fill out the survey forms and mail them to the collection agency - they are almost always included in the early estimates. On the other hand, bad respondents are often too late to be included in the early estimates and too often must be reminded through telephone follow-up to complete the survey form.

This paper examines the feasibility of conducting surveys through the use of facsimile machines, personal computers and intelligent character recognition (ICR) software - Intelligent character recognition is sometimes used to refer to software which can recognize and convert typed or hand printed characters to ASCII. We will try to determine if these technologies have progressed to a level where they are competitive with the traditional method of data collection; i.e., "mail-out/mail-back" and clerical keying. We examine the state of the art of the ICR software and what does it cost to use this technology.

Technology Definition: For the purpose of this study, PAPERLESS FAX IMAGE REPORTING SYSTEM (PFIRS) is defined as the process of receiving questionnaires, as an image file, via facsimile transmission, and using computer software to read the survey image files. The system must: 1) be available, over the public telephone network, twenty-four hours a day, 2) be capable of determining the survey and respondent to which the questionnaire corresponds, 3) have the ability to display the document on a computer terminal for keying, and 4) have Intelligent Character Recognition (ICR) capability. Additionally, the system should have the capability of FAXing the survey instrument, an advance notice, or a non-response reminder to the respondent. Finally, the system should be able to perform all the above system requirements in a completely paperless environment.

In order to determine if a FAX imaging and ICR system was feasible, we conducted an initial technical assessment. This assessment examined the following:

1. <u>Stage of Development, Stability of Technology</u>. Is this an emerging,

untested technology, a well established one, or something in between?

2. **Range of Potential Applications.**

3. **Initial Investment Required.**

4. **Ease or Difficulty of Setup or Authoring.** How much work is necessary to prepare it for use in a specific survey at the present time?

5. **Respondent Acceptance.**

6. **User Training Required.**

7. **Effect on Coverage.** What are the limitations for population coverage that may be induced by the technology?

8. **Effects on Response Rate.** What are the effects of use of the technology on survey response rates?

9. **Confidentiality and Security.**

10. **Editing and Related Capabilities.**

11. **Effects on Estimates.**

12. **Effects on Timeliness.** Does the technology increase or decrease the timeliness of reporting or of the preparation of the estimate?

We will discuss some of our findings related to the above issues and briefly outline the unresolved issues needing further research.

Stage of Development, Stability of Technology.

In assessing the current state of the technology, two areas are considered:
1) the facsimile technology used to send and receive the forms and
2) the ICR technology used to convert the received FAX image files to ASCII.

The facsimile technology is well established. There are numerous companies manufacturing facsimile machines. "Currently, there are at least 60 U.S. based companies manufacturing fax modems" (Skipinker, 1992). In recent years, we have seen the introduction of the less costly FAX board. These boards enable personal computers to send and receive facsimile images. These images received via facsimile transmission are usually stored in one of several image file formats. There are a number of quality control precautions built into FAX modems/boards which help to ensure data quality. "In late 1991, modem companies started delivering products that conformed to a revised signaling standard called V.32bis. Modems following this standard offer faster signaling than V.32 models and smoother fallback to greater choice of speeds .... The newer V.32bis modems more extensively analyze the connection to determine immediately the best usable signaling speed ..."(Derfler, 1992). Perhaps the best evidence that this is an established technology is the fact that one can purchase a FAX board for under one hundred dollars.

In May of 1992, the Census Bureau sponsored an ICR Conference at NIST (National Institute for Standards and Technology). The results of the conference suggest that the software to recognize hand-printed alphanumeric text is usable but still needs development. The NIST results suggest that these products are at least 95% accurate at recognizing hand-printed digits, at least 90% accurate at recognizing upper case hand-printed letters, and 84% at recognizing lower case hand-printed letters. It should be noted that the NIST images were obtained by scanning, and that these results might not be duplicated when performed on FAXed images.

Range of Potential Applications.

The literature is replete with numerous applications of either facsimile or ICR technology. Some of these involve data capture and deposit, advertising/marketing, polling, and inventory control. The Bureau of Labor Statistics (BLS) has recently used facsimile technology as a substitute for telephone and mail to notify Current Employment Statistics Survey

(CES) establishments of the approaching reporting deadline (Rosen, 1992).

The Census Bureau currently receives some survey forms via FAX. For example, Industry Division's M3 survey with approximately 3000 respondents, receives approximately 900 reports by FAX.

The Wyoming Department of Revenue recently used ICR technology to develop a data capture and deposit system (DCSD). This system was used to recognize hand-printed state tax forms and store the data into a database. Typically, one sets a lower bound for the level of confidence for each character which is to be recognized by the ICR. If the ICR's level of confidence for a data item falls below this lower bound, the offending item is flagged and displayed on a computer terminal for clerical verification. The Wyoming Department of Revenue intends to integrate the data capture and deposit system with a FAX server. "The fax server will accommodate inbound tax returns filed electronically in the future. Outbound fax transmission provides a general purpose, low cost communication vehicle to other departments and to remote users" (Farmer, 1992). Several banks use ICR software to perform check processing. The postal system uses ICR software produced by AEG, Inc., to sort mail.

Each month, *PC World* magazine conducts a survey of its readers. Respondents are asked to indicate whether they read a specific article and the degree to which they found it useful. After answering the questionnaire, respondents are asked to FAX it back to *PC World.* The FAXed-in image is then interpreted by ICR software and the responses are stored in a database.

Initial Investment Required

The character recognition technology has now progressed to the point where one can purchase pre-customized packages. For example, Cardiff, Inc. now sells a product, Teleform, which can be used to interpret FAXed images and store the data into a database. The one caveat is that the images must be of forms created by Teleform. The price ranges from $1000 to $4500.

One can also purchase ICR development tool kits in order to customize an application. If one chooses this route then software to perform forms removal, remove background noise, and perform de-skewing should also be considered. Dropout ink should also be considered.

In addition to software, one must also consider the cost of hardware. Keep in mind that image files can require a lot of storage space - one eight and a half by eleven inch sheet of paper can require up to one megabyte of storage space, depending upon the image file format. Also, most good character recognition software are CPU intensive. And finally, a good monitor is recommended for viewing the images.

Effect on Coverage.

What are the limitations for population coverage that may be induced by the technology?
The technology is clearly limited to respondents having FAX capability.

Confidentiality and Security.

What special problems (if any) does this technology pose for protection of respondent confidentiality and/or security of the Census Bureau's computers, programs, or data files?

For the respondent: In similar systems, the surveying agency assumes responsibility for the data only after it has reached the destination. This is really no different from receiving data via mail.

For the Bureau: The FAX server(the machine receiving the facsimile transmissions) should not be part of the Bureau's computer network. Moreover, the system will only be receiving FAXed images; there is no modem involved. Therefore, there is no way a respondent can interact with either the computer's operating system or stored files.

Editing and Related Capabilities.

Several of the ICR software packages examined allow for the validation of data items against databases. Data items can also be stored directly to databases and thus edited by auxiliary programs.

OUTSTANDING ISSUES:

Our assessment of facsimile and ICR technologies indicates that a paperless fax image reporting system should be feasible, but we could find no definitive published example. Therefore, our intention now is to build a prototype PFIRS system. We will first test "off the shelf" ICR software specifically designed for interpreting FAXed images. However, we will also procure integration software, i.e., ICR development tool kits, in case the package cannot be customized to meet our needs.

One of the objectives of our research is to determine if our current survey forms can be used with an image recognition system. Currently, our survey forms contain horizontal response areas in which respondents are supposed to enter the data. We believe that a potential problem with this type of form layout is that respondents will print characters too close to one another, thereby inhibiting proper recognition. As is often the case, some respondents will write script in this type of form layout. Currently, no commercially available software exists which is capable of cursive recognition. It is possible that the response areas on the survey forms will have to be laid out in blocked form. We believe that this format not only helps to prevent cursive writing, but it also encourages true print form.

A second major concern with regard to form design is storage. We would like to archive the images files received from the respondents. We will consider forms removal techniques. These techniques can be used to remove horizontal and vertical lines for the images. The more efficient image formats will store only the non-white information on a sheet of paper. Of course, we will also consider the use of dropout ink which does not appear on the faxed in image files.

Another concern is form skewing and background noise. A form carelessly placed in a fax machine will result in a curved or bent image on the receiving end. Finally, a poor telephone connection may result in an image with background noise - these often appear as black dots randomly scattered on the image. We will integrate front-end software which will perform skew correction and background noise removal. This cleansing of the image will, of course, occur before it is dispatched to the recognition software for interpretation.

While forms removal software will do much of the work of clerical workers, it will not totally replace them. We believe that in a complete image recognition system, it must be possible to display the form's image and compare the respondent entries on the image with the ICR software's interpretation of those entries. This leads us into the realm of the human computer interface. We will be exploring several options to determine the correct screen layout for the clerical review phase of the image processing system.

CONCLUSION: "Conventional fax has become ubiquitous. Like the copier a decade ago, it seems every business has conventional fax capability. Fast, cheap, convenient and efficient, fax has become a commonly accepted mode of business-to-business communication" (Radding, 1992). If the respondent has facsimile capability and if the ICR works well enough, PFIRS appears to be an inexpensive paperless alternative, or adjunct, to traditional data collection (MAIL, CATI - computer assisted telephone interviewing, clerical transcription of data). It has distinctive advantages for both the Bureau and the respondent. Some of these include:

1. Reduction in clerical keying without sacrificing accuracy. Several of the ICR software reviewed in our technical assessment allows the user to set acceptance levels. For example, each character converted from hand print to ASCII has an accompanying "level of confidence" which indicates the

confidence that the ICR system has that the conversion is correct(this confidence is obtained by comparing the result to those obtained during the training phase of the neural network software). The user may choose to review all data items, or only those which do not satisfy the confidence level constraint. Data items can also be validated against databases or data dictionaries.

2. We are hopeful that the reduction of our reliance on the postal service coupled with a significant reduction of clerical keying, will result in more forms being included in the initial estimates.

3. The software can be integrated with databases on line so that verifications and edits are performed as the forms are processed. This should also result in improved initial estimates.

References:

Derfler, Frank J. "Maximum Modems, 14,400 bps and Rising." PC Magazine, March 17, 1992. Vol 11, N 5, pp 285-302.

Farmer, Gerald. "Automated Data Capture & Deposit System Utilizes OCR for Hand-printed Tax Forms," *IMC Journal*, pp 23-26.

Radding, Alan. "The Fax of Life: Computer-Integrated Fax Changes the Way Organizations do Business." *Midrange Systems*, February 18, 1992. Vol 5, N4, p 13.

Rosen, Richard J. and Clayton, Richard L. "An Operational Test of FAX for DATA Collection." Bureau of Labor Statistics, 1992.

Skapinker, Mark. "Computer FAX Finally Overcomes its Limitations." *Computing Canada*, Vol. 18, N2, p51. January 20, 1991.