Utilization of Document Imaging Technology by the 1996 Canadian Census of Agriculture

Mel Jones
Ivan Green

Statistics Canada

ABSTRACT: Processing of the Canadian Census of Agriculture relies heavily upon access to its questionnaires. The documents are required within many stages of collection and processing to ensure delivery of quality products to agricultural clients. Historically, access to the physical documents has proven problematic in terms of timely delivery, concurrent access, management, tracking, storage and cost. As early as 1984, Census management recognized the potential of document imaging technology to improve data quality and timeliness. However, in the 1980s it took little effort to recognize that the technology of the day was not sufficiently mature or affordable for the Census. In 1996, document imaging became a successful reality for the Canadian Census of Agriculture. This paper describes the experiences, the issues and the expectations of the many different players involved in the implementation.

Introduction

Implementing document imaging technology for the Canadian Census of Agriculture involved significant expense and risk taking. In the early stages of the project, healthy discussions took place on conventional methods and competing technologies. The first section explains the issues embedded in the discussions and sheds light on the direction chosen. It also contrasts the drawbacks of the historical manual methods with the expected benefits of document imaging.

The technical challenges associated with implementation of document imaging were many. The second section touches briefly on several of the challenges and explains how key technical issues of document imaging were identified and satisfied. It closes with observations on some obvious and not-so-obvious reasons why the introduction of the technology succeeded.

Section 3 contains a management perspective on the decision to use imaging technology and offers some possibilities for its continued use and evolution in Statistics Canada.

1. Why Imaging?

In accordance with agency policy, release dates for publications are made available in advance to the public. When declared, the release dates do not change. Consequently, statistical production adheres to very strict deadlines to meet release dates.

Time and fiscal constraints, in addition to the need to maintain or improve quality, have fueled a drive to become more efficient. As a result, increased emphasis is being placed on automation and innovation. Expensive manual processes are prime targets.

Prime targets for the Canadian Census of Agriculture were (a) data capture and (b) the delivery of physical questionnaires throughout the production workflow. Feasibility studies to automate were conducted for both.

For data capture, an investigation of intelligent character recognition (ICR) technology was carried out. At the time, the maturity of ICR technology for Census purposes was in question and there was a lack of time to test respondent reaction. To use it would have significantly increased respondent burden. This is because a change from unconstrained handwritten numeric responses to ICR-friendly, constrained numerics would have been required. Since more than 90 percent of the information sought by the Census is numeric, a complete redesign of the questionnaire would likely have been necessary. After the feasibility study, the efficacy and cost-effectiveness of recognition software was also in question.

The research into imaging and ICR technology lead to a very interesting and challenging situation. In Canadian political terms you would say the project management team fell into 3 camps: the conservative, the progressive-conservative and the progressive. The conservatives felt that implementing imaging by itself was too high risk because:

- We were 2 years away from the Census. For a quinquennial project, it was too high a risk to test and implement imaging in that time frame.
- The cost estimates varied widely from "affordable" to "way out of our league".
- The department had no experience with imaging and ICR technology. It would be safer to prove the technology on smaller and more frequent surveys.
- The staff would be totally dependent on machines. If the system went down, staff were out of work.
- The project had many other challenges including:
    - the addition of a new Progress of Seeding Survey, a CATI follow-up survey to over 40 percent of Census respondents;
    - the complete re-engineering of processes and migration from the main frame to mid-range technology; and
    - a new system and methodology for the register of farms.

The conservatives understood the value and benefits of imaging but felt the risks were too high for the 1996 Census of Agriculture. The traditional manual method (no imaging) was safest and proven.

The progressive camp believed the greatest savings could be realized by implementing imaging and ICR together. They believed imaging without ICR was not cost effective. They were not only convinced that imaging could be done, but that ICR technology was then sufficiently mature for Census purposes.

The project manager's original proposal was to implement imaging without ICR. The progressive-conservative group supported it. Risks associated with ICR were unacceptable given the time frame faced by the project. This camp, however, felt imaging was feasible with a manageable amount of risk. While they had heard of many disastrous imaging projects, they had also seen enough working applications in other locations to convince them that the risks were worth the benefits.

There was no right or wrong in the debate. Each group had valid reasons for their stand. Once the decision was made to implement imaging without ICR, the decision was fully supported by the entire project team. Delivering electronic images of questionnaires to the desktops of editors and validators was now the goal.

If a manual questionnaire delivery system had been implemented for 1996, it would have suffered from the same shortcomings as its predecessors:

- *Timeliness:* Delivery of questionnaires to production staff would take hours, and in some instances, days.
- *Contention:* This is another timeliness issue. Often, validators and editors would require concurrent access to a questionnaire from a large farming operation. Large farming operations are the top contributors to estimates for many Census variables. Delivery would be conducted in a serial manner because copies could not be made. Having duplicate copies was considered a security risk and was not permitted. To keep track of documents, the questionnaires were filed and pulled again between each operation or request.
- *Tracking:* Maintaining an awareness of the location of each document would require development of an automated tracking system. A manual record of receipt via barcode readers would be needed at each stage of production. This type of system would impose a serial movement of questionnaires (and data) through the processes.
- *Complex storage:* An elaborate system of folders, boxes and shelving would be required to support high-volume retrieval and storage of documents.
- *High costs:* Staffing, shelving, storage space, security and development of a questionnaire distribution and tracking system all contributed to an expensive, comparatively slow manual system.
- *Frustration:* Validators had to make lists on paper of needed questionnaires and then submit these to the library. In a matter of hours they normally would get their request back. However, many would be missing due to contention, filing errors, etc. Several requests would have to be made before all documents were located.
- *Filing errors:* Filing and re-filing resulted in errors. Periodically, the library operation would be shut down to do a systematic and complete check of the library.

Intuitively, a questionnaire imaging system addressed the shortcomings of its manual counterpart:

- *Timeliness:* Delivery of questionnaire images to production staff would take seconds rather than hours or days.
- *Contention:* All users could examine the same document concurrently. This would create opportunities for *parallel* approaches to processing.
- *Tracking:* No manual questionnaire tracking system was necessary throughout processing workflows. In addition, people were spared the effort of documenting receipt and release of questionnaires.
- *Complex storage:* A temporary storage would still be required. However, the system of storage would be comparatively simpler.
- *High costs:* Since imaging had never been conducted on a large scale before at Statistics Canada, there was no benchmark with which to compare the Census of Agriculture. As a consequence, it was not clear as to whether imaging was more or less expensive than a manual questionnaire delivery system.

Management believed that the potential benefits warranted a closer look. In addition to those benefits listed above, management expected the following:

- *no need to microfilm* the documents after processing;
- *reduction in manual operations*, particularly the library staff;
- *immediate access to documents* would make staff more efficient as they could find a problem and solve it immediately, no need to make a list and then find something else to do until the documents arrive;
- *reduction of paper handling* by all clerical and validation staff, as delivery of electronic images would eliminate much of the need to handle paper;
- *remove frustration* caused by delays in receiving or waiting for questionnaires and missing questionnaires;
- *job enrichment*, as staff would find previous mundane jobs satisfying and challenging;
- *immediate access to problem questionnaires* meant staff could rely more on information than intuition (previously, in order to avoid waiting for the paper document some staff would guess at the problem and the solution); and
- *immediate access* would allow staff to look at more problems and thereby improve the quality of the estimates, most noticeably at the lower geographic levels.

2. Imaging Issues

When the investigation of document imaging began, answers to the following questions were sought:

- What is document imaging?
- What components of imaging are necessary to meet our requirements?
- Will it work in our technical environment?
- What is a rough estimate of the costs?
- What issues must be addressed for successful implementation?

As a first step towards answering these questions, Census representatives began discussions with the private sector. Industry was provided statements of requirements, volumes and samples of the 1991 questionnaire which was considered to be representative of the (as yet to be developed) 1996 version. In return, industry representatives provided cost estimates and descriptions of potential solutions. Estimates ranged from C$500,000 to almost C$3,000,000. The lower estimates were affordable, the upper were not.

The solutions delivered with the estimates were somewhat confusing. They addressed some problems that were not clearly applicable to the Census and omitted others that possibly were. However, despite some minor confusion, Census representatives began building a clearer picture of what the issues were. Some emerging (and very expensive) issues which concerned the Census were:

- Indexing,
- Quality control,
- Throughput of scanners using legal-sized, 70m Vista paper (the Census questionnaire),
- Required adjustments to the layout of the questionnaires to make them*image-friendly*, and
- The quality of the images.

To build a sufficiently complete list of issues, seminars on document imaging at Association of Information and Image Management (AIIM) trade shows were attended and a visit with imaging researchers at the United States Bureau of the Census (USBC) was made.

USBC representatives brought to light the critical issue of *document integrity*. USBC questionnaires are multiple-page booklets. Scanners deal with one sheet of paper at a time so imaging a multi-page document may require cutting to allow access to the individual pages. Of primary concern to USBC researchers was the potential for pages of a multi-page document to be lost or become out of sequence during movement to or from the scanner; i.e. loss of document integrity. At the time, research of page-turners (rather than cutters) was being undertaken by USBC. Since the Census of Agriculture's questionnaire is multi-page, it was believed that this was an issue for Statistics Canada as well. Given the relatively small volume (280,000) of questionnaires for the Census, page-turners did not appear to be a cost-effective solution.

One of the first solutions to come to mind was to print a unique questionnaire identifier on each page of each questionnaire. Discussions were held with internal experts and with large print shops. The solution was indeed feasible, however, not cost-effective. Estimates provided were close to C$400,000, adding approximately C$1.43 to the cost of printing each questionnaire and more than tripling the budget for printing. The chance of error during sequencing process still existed. A second solution considered was to have the scanner emboss a serial number on the bottom of each page that it scanned. This was a suitable solution for questionnaires dropped *after* they had been scanned; it still did not address those dropped on their way to the scanner. More work would have to be done.

Investigation of document imaging continued. Trade shows illuminated the following additional issues:

- Size and quality of monitors,
- Video cards and functionality,
- Network capacity,
- Storage capacity,
- Scanner functionality,
- Technical support, and
- Integration with in-house applications.

At this point, a reasonably complete list had been compiled, but it still was not clear how it all applied to the Census of Agriculture. There was a concern that the Census application might have some abnormal imaging issues to address. For example, most scanners are rated on single sheet, letter-sized, 30-weight paper. The Census uses 16-page booklets of legal-sized 70m Vista paper. Management agreed to fund a small pilot (C$20,000) to clarify the Census issues.

A local firm was contracted to set up a basic imaging system on-site. They worked closely with internal systems and methodology experts. The objectives were:

- to investigate workflow methods and procedures to fully operationalize imaging in terms of questionnaire receipt, preparation, scanning, reassembly, storage, retrieval and quality control;
- to become familiar with industry standard image hardware and software capabilities; and
- to become familiar with performance of the technology with respect to Census documents.

The pilot system was composed of:

- two different scanners (Fujitsu and Bell & Howell),
- a Hewlett-Packard server running UNIX and LAN-Manager 2.1,
- four PCs with 2 MB video cards,
- 17 inch and 21 inch color and grey-scale monitors,
- TCP/IP,
- Ethernet 10 Base T, UTP subnet,
- Watermark TIF image viewers, and
- a sample of 2,000 16-page legal-size questionnaires from the 1991 Census.

The pilot lasted three days and yielded the following results.

- Political acceptance of the technology increased. Observation of questionnaires being scanned, the quality of images and the functionality of image viewers boosted confidence and reduced skepticism.
- Due to Statistics Canada's adherence to interface standards, there were no unexpected surprises arising from the technical infrastructure.
- The observed throughput of the scanners was considerably less than rated, but was still impressive.
- In the Census experience, equivalent models of properly serviced scanners from competing manufacturers are not equivalent regarding throughput. One of the scanners performed reliably, having relatively few double-feeds, skews and jams. The other scanner was very poor, literally destroying questionnaires; its use in the pilot was quickly discontinued.
- The functionality of scanners was quite extensive in terms of reading barcodes, indexing, counting, etc.
- The number of scanners of this type and the staffing required for Census volumes was determined.
- The quality of the images was acceptable (200 dpi).
- The average size of the images (produced by the pilot scanners) was determined. This information would be useful in simulating network traffic and in determining image storage requirements.
- The weight and color of the paper used for the questionnaires (Vista 70m opaque) was acceptable, as no markings from one side of a sheet of paper bled through to the other.
- The layout of the 1991 questionnaire would be unacceptable for imaging. However, adjustments would be minor and would have no impact on the respondent or on printing costs for 1996. No testing of respondent reaction to the image-friendly questionnaire was necessary.
- The functionality, look and feel of a Watermark viewer would meet Census requirements. This would allow a crude estimate of viewer costs and serve as a reference point for specification.
- Document integrity was not an issue. Concern existed over cut documents being dropped and pages lost or misplaced; i.e. loss of integrity. Tests demonstrated that batches of cut questionnaires tended to stay together when dropped. This was due to static electricity and folds in the pages. Nevertheless, the risk of lost or misplaced pages still remained. Measures were taken to further reduce the risk. They included keeping the batches small (30 or less documents), storing batches in envelopes, minimizing time out of the envelopes, reducing the

distance between cutter and scanner, using handwriting style and ink color to reassemble, and as a last measure of insurance, using either the images or the data in the structured database to reassemble dropped documents.

- The less expensive 17 inch color monitors and video cards were acceptable (at 1024×768) for at least half of the Census production activities. The quality of the image was often*better* than the original, due to the fact that images can be enlarged. The grey-scale monitors provided a clearer display, but not significantly. The color monitors were considered easier to deploy after production to other activities.

The results of the pilot test paved the way towards implementation. What remained was:

- Quality control strategy,
- Indexing strategy,
- Determine network requirements,
- Determine integration requirements,
- Determine scanner requirements,
- Determine viewer requirements,
- Selection and acquisition of monitors and video cards,
- Acquire storage technology,
- Develop the workflow and control systems for the scanning operation,
- Develop an RFP, and
- Implementation.

Throughout all of the remaining activities, the contributions made by representatives from the Agency's Informatics User Services Division, the Mid-Range Support Group, Business Survey Methods Division and the System Development Division proved to be of critical importance.

2.1 Bandwidth

A major concern regarding image production loads was the network. A few years before the project, the Agency had rewired it premises to 10 Base T UTP. Despite this, the wiring had proven to be vulnerable to movements of large files. Assurance that it would support the concurrent activities of 100 end users and a high volume scanning operation was needed. The fact that no benchmark existed within the Agency for imaging volumes posed a problem. Fortunately, key task managers from the previous Census were still available and provided information about the volume of questionnaires drawn on a daily basis using the manual system. It was anticipated that imaging volumes would exceed that of the hardcopy documents. For testing purposes, the volume was tripled to represent user traffic. It was added to the expected volume of scanning traffic (total volume of questionnaires divided by the total number of days available to scan) to represent the total daily imaging load on the network. The total load was divided by 6.5 hours (average level of activity in 7.5 hour day) to yield the average load at any point in time during production. Using this formula and the information regarding image sizes obtained from the pilot, a series of network tests were conducted.

Experts from the Agency's Communications Services Center (CSC) were called to assist in the tests. After consultation, CSC provided advice on how to effectively conduct tests to yield the most meaningful results. Simulating the megabits per second demand, i.e. a workstation pulling and pushing images to and from a server over the network, was insufficient to represent 100 users. Simulation of the *distribution* of the demand was also important. If only one workstation pulled and another pushed,

real-life network collisions were not likely to occur. Many workstations had to be employed if the tests were to be meaningful.

Prior to conducting the tests, CSC configured an isolated subnet. This was done to eliminate noise so that the loads being measured were attributable to movement of images. Twenty workstations were used. On each was a program that either copied a file to or from a server on a specified time interval. CSC used network analyzers called "sniffers" to observe the traffic.

Based on CSC recommendations, CSC acquired an Ethernet Routed Switch with 100 MB capacity to support production loads. Representatives from the Technical Support Group installed the switch, isolating the Census users from the rest of the Agency. The purpose was two-fold: (1) to protect the imaging subnet from performance degradation due to unanticipated fluctuations in external LAN traffic, and (2) to protect external LANs from the Census imaging loads.

Census users were organized into three workgroups on the Ethernet switch with a 10 MB connection. The server was linked to the switch by a 100 MB fiber optic cable. Census funded acquisition of a 100 MB Fiber Direct Data Interchange (FDDI) card to interface the server with the network. The Mid-Range Support Group investigated solutions, made a recommendation, and after consultation, acquired and installed the FDDI card.

After the *image subnet* was implemented, CSC returned to conduct further tests to ensure that expectations had been met. They had been.

2.2 Monitors

Since the imaging system exists to serve the end-user and most often the only awareness of the imaging system is based on what is on the screen, the monitor is arguably the most critical component of an imaging system. It is certainly one of the most expensive. During the pilot test, no significant difference in the appearance of Census images was observed between the 17 inch color and the 19 inch grey-scale monitors. Since the monitors were to be used for more than imaging applications, despite the higher cost, it was decided that color monitors would be acquired. What remained was to determine (a) the size or sizes of monitors that meet the Census needs and (b) the video card requirements.

Some stages of the Census production workflow require concurrent observation of two or more legal-sized pages. It was not known if a 17 inch monitor with an appropriately configured video card would or would not be sufficient. The cost difference between a 17 inch and a 21 inch is very significant. It had to be determined if the expensive larger monitors were actually required.

At the time, there were several 17 inch monitors being used by Census staff that could be used for an imaging evaluation. However, larger monitors were not available. Representatives from the private sector were contacted for some 21 inch loaners.

Technically oriented members of the Census production staff volunteered to evaluate monitors and video cards. Large and mid-sized monitors were rotated amongst the group for a period of three weeks. It was determined that in spite of *virtual desktop* capability and a resolution of 1280×1024, the 17 inch monitors were insufficient to meet all the Census requirements. Virtual desktop is a feature of video cards. It allows the user to specify desktops that are larger than the physical dimensions of a monitor's display surface. The areas of the virtual desktop that are not visible can be brought into view

by movement of the mouse. The virtual desktop was considered essential for some of the activities to be performed on the 21 inch monitors. Based on the research, the quantity of medium and large monitors was determined. The budget was next.

The pricing, physical dimensions and performance of monitors vary considerably. For example, it was necessary to ensure that the 21 inch monitors would comfortably fit on the furniture; some did not. Perhaps more importantly, the quality of the display in terms of sharpness, contrast, brightness and consistency of image on the entire display surface varied considerably among comparatively configured and priced monitors.

The Agency maintains standing offers (SO) with vendors to facilitate acquisition of informatics technology. On the SO list were offerings from several vendors. It was in the project's best interest to purchase from the SO if any of the monitors on the list met the requirements. A representative from the Agency's Workstation Support Group provided loaners of all the 21 inch monitor types on the list. The monitors were simultaneously driven by the same workstation and displayed the same image. Prices ranged from C$900 to C$3,000. From a distance, the least expensive monitor looked good. However, closer inspection revealed blurring and fading of colors at the edges of the display. Some flicker at a resolution of 1024×1280 was also noticed. The higher priced monitors performed very well but were more expensive and lower in quality than one of the models previously evaluated by Census staff. Arrangements were made with the vendor to have their product added to the SO. (The desired video cards were already there.) These monitors were included in a purchase of 120 Pentiums destined for production.

During production, it was noticed that many of the end-users did not take advantage of the virtual desktop feature of the video cards simply because they were not comfortable with it. In addition, many users did not use the higher resolutions supported by the monitors and video cards. Many reverted to a resolution roughly equivalent to that on their former 14 inch SVGA monitors. Perhaps the clear, small text supported by the technology is too small to read, especially for people with visual impairments.

2.3 Quality Control

Many imaging systems have workstations and staff responsible for the quality of the images made available to production workflow. QC measures can significantly affect the cost of document imaging, often requiring expensive equipment, consuming salary and potentially introducing a bottleneck.

The bottleneck was of particular concern. Most workflows involving document images are image driven. The Census workflow was driven by data. This means that when an editor or analyst had access to the captured data on the structured database, they were ready to begin work. If the corresponding questionnaire image had not yet been delivered, they would either continue without its benefit (defeating the purpose) or wait until it was available. For the Census, it was imperative that images were delivered either *before* or *at the same time* that structured data was. To inspect (in a timely manner) the daily volumes of images generated by the scanning operation without causing delays in production workflow would have required an expensive QC operation.

Discussions were held with Census management, key staff, methodologists and systems personnel to look for ways to circumvent the problem. As long as the page was legible, it was not of concern to the user if the rest of the document was corrupt. Most often, end-users only looked at a portion of one page of each 16-page document they requested. A methodologist, reviewing information (statistics on double-feeds, skews, tears and folds) obtained from scanner evaluations, determined the expected

number of times an end-user would encounter an unacceptable image and request a re-scan. They were negligible. This was due to the consistently high quality of the images generated by the scanner. As a consequence of this investigation, it was decided that QC operations would be performed after, rather than during, the time critical production period. This decision was made easier due to the fact that to request a re-scan required the end-user to simply click on a *re-scan button*.

As a result of some creative thinking and closely focusing on the need rather than implementing conventional technical solutions, the amount and cost of QC was significantly reduced. No additional workstations were required and staff was tasked with QC during their slack time.

2.4 Indexing

As with other types of database entities, one or more keys are associated with an image to support timely retrieval by the end-user. Some applications employ optical character recognition (OCR) to automatically extract identifying information from an image. The information would then be used to build indexes.

Since the identifying questionnaire information was in unconstrained handwritten form (rather than typewritten or printer generated), OCR was not a feasible solution for the Canadian Census of Agriculture. ICR had already been ruled out. An alternative was to build an interactive indexing application. The implementation would require staff to view selected pages from each questionnaire image and data capture the identifying information. This data would then be used to build indexes to the imaged documents. Two drawbacks existed: (1) index stations would increase the cost of imaging, and (2) another potential bottleneck would be introduced prior to the production workflow.

Fortunately, the necessary indexes could be obtained elsewhere and at no additional cost. All appropriate identifiers referenced the database, which housed the data-captured questionnaire information. A link was required between the images and the database. Representatives were called in from the Agency's System Development Division to seek a solution.

It was known that scanners could read barcode labels and automatically register the codes as indexes to images. Based on data capture plans, the barcodes were to be affixed to the questionnaires at the time of keying and were to be captured in addition to the respondent data. The codes would be used as unique identifiers on the database.

Given this scenario, systems people proposed that a program be written to (a) access new images on the image base to obtain a barcode number and the image base internal identifier, and (b) using the barcode, find the matching record on the database and store the image base internal identifier within it. This would permit end-users to access a database record using any of the predefined keys and then select the corresponding image from the image base using the internal image identifier. This was an effective, yet relatively uncomplicated solution.

2.5 The Shape of the System

What was once very cloudy regarding the imaging solution was becoming clear. Since QC applications, indexing applications and embossing of questionnaires were not required, the essential software components that appeared to remain were:

- Scanner management software,
- Image base management,
- Image viewers, and
- Bridges between the image system and the Census application (integration).

The hardware components were:

- Scanner(s),
- Server,
- Storage device(s),
- Network infrastructure, and
- Workstations, monitors and video cards.

The paper component (the questionnaire) had been dealt with in terms of color and had been made *image-friendly*.

2.5.1 Scanner Management Software

Much had been learned about scanner functionality and performance through the pilot test, trades shows and through discussions with industry. However, the evaluations using Census paper had not been applied to an industry leader, Kodak. This appeared to be a significant omission because Kodak was just as probable of being in the winning RFP as the other scanner manufacturers. Additionally, (other than using a barcode as a key) the critical details of how a scanner was to manage Census documents had not been addressed.

Given the higher-end capacity of some of Kodak's models relative to the scanners in the pilot, fewer staff might be required in the scanning operation. A meeting between a Kodak representative and a Census manager tasked with operationalizing the scanning operation was held. An on-site loan of an ImageLink 500 was arranged. Census representatives, a methodologist, microcomputer support staff, systems people and Kodak representatives worked closely for a month. In addition to gaining a familiarization of the functionality and performance characteristics of a Kodak, comparisons could be made to offerings by other manufacturers used in the pilot test. A sense of which scanner technology would best suit Census needs was developing.

Tests were designed to address as many aspects of scanning for the Census as could be imagined. The operations manager, her assistant and a methodologist focussed on the throughput of the scanner and the quality of images, determined whether a white light was sufficient to scan both colors of the questionnaire, and learned what was involved in operating and maintaining the scanner.

Critical outputs from this exercise were:

- scanner functional specifications to be included in the RFP,
- benchmark figures (occurrence of double-feeds, skewed images, jams),
- average sizes of acceptable Kodak images (these were consistent in size with images produced in the pilot test),
- a stronger sense of the staffing issues,
- a discovery that the Kodak transport mechanism worked better with half rather than full batches (15 versus 30) of Census documents, and

- a discovery that barcode labels must be very precise or the scanner will not recognize them. All barcode labels are not of equal quality. If the *hands-on* evaluation had not taken place, this critical discovery might have been made at a potentially crippling time.

Following this evaluation, scanning management began addressing the considerable challenges associated with operationalizing the scanning activities. This involved designing a workflow which dealt with receipt of documents, sorting, scanning and eventual storage. An automated system to track the movement of documents throughout the normal workflow and through re-scan requests was designed and specified. An Agency systems representative built it (Visual Basic), participated in the test activities and installed it. The scanning operation functioned very efficiently and very reliably in production.

During the Kodak evaluation, a senior Census Task Manager began managing the development and delivery of the RFP. He recruited the operations manager, her assistant, the Census technical liaison, a representative from Mid-Range Computers and a representative from the System Development Division. The contents of the RFP were defined to be:

- Background information about the Census,
- Information on the technical infrastructure (server technology, OS, network protocol, desktop configuration, desktop OS, etc.),
- Scanner functionality and performance requirements,
- Training requirements,
- Integration requirements,
- Technical constraints such as image sizes,
- Image viewer functional requirements,
- Support requirements,
- Acceptance criteria, and
- Time constraints.

Issuing an RFP in the Canadian Federal Public Service is subject to many rules and regulations. Although necessary, the result is that a typical RFP can consume up to six months from delivery to Public Works to eventual declaration of a winner. Time was becoming a critical factor for the project as production was less than eight months away. It is believed that the RFP process could have placed the project in jeopardy had the manager not kept in constant contact with Public Works, pressuring them to keep the Census RFP as a high priority.

2.5.2 Image Viewers

Options existed in the market; viewers could be acquired or custom ones built. Some industry representatives advised that considerable savings could be achieved if a custom one were to be built. Others advised against it because of proprietary issues. It was not known whether the winner of the RFP would build a new viewer or utilize an existing one. Either way, the functional requirements had to be determined and included in the RFP.

The determination of requirements turned out to be relatively simple. A Watermark image viewer was obtained and made available to selected Census staff. After a demonstration and some time to evaluate the software, a walk-through was held. During the exchange, the individuals simply identified what they felt was required, what should be modified and then prioritized it. They also indicated what they felt was not required. The results were then documented and included in the RFP. This ensured that

the viewer requirements would be met and whether off-the-shelf or custom-built technology was to be delivered by the contractor.

2.5.3 Storage Requirements

This was relatively easy to do. After examining samples of images obtained in the pilot and in the evaluation of the Kodak, an average image size of approximately 600K was determined. The size was multiplied by the number of questionnaires (along with a 25 percent insurance factor) to derive the total storage requirement.

Some risk was involved here, however. This was due to the fact that the RFP for storage technology had to be let before a winner of the imaging RFP was declared. The size of the image had a direct impact on three critical items:

- *Image quality:* The larger the image file, the better the quality. Care had to be taken that the images were acceptable.
- *Storage requirements:* When dealing with large image bases (4.8 million images) subtle changes in image sizes can significantly affect storage requirements. Adequate storage had to be available prior to production.
- *Performance:* The larger an image, the greater the load on the network, the server and the workstation. All the preparatory performance tests had been done with an anticipated image size. A significant increase in image size may have had adverse affects.

2.5.4 Performance

A critical concern for the Census production managers was the speed at which a requested image is delivered to the desktop. It was potentially the difference between success and failure in Census imaging.

Minimizing delivery times was a challenge. This was due to the many layers of technology which delivered images, and competition for computing and network capacity by other processes.

Performance issues included:

- The desktop
  - Clock speed
  - RAM
  - Cache
  - Swap space
  - OS (16-32 bit)
- The monitor and video card
- The Network
  - Capacity (10-100 MB)
  - Protocol (TCP/IP, packet sizes, etc.)
  - Topology (isolated, not isolated)
- Images
  - Size of images
  - Size of image base
  - Database technology
- Server
  - RAM
  - OS
  - Dedicated or supporting other application
- Storage devices
  - RAID caching
  - Jukeboxes
  - Optical medium density
- Imaging application
  - Viewers
  - Image Base Management
  - Indexes, etc.

As is evident, holding the contractor to specified performance levels in production was not possible. This was due to the many environmental factors outside his control. However, it was still essential to determine that the performance of contractor's solution would not jeopardize the project.

A performance test was designed, described in the acceptance criteria of the RFP and implemented. In an isolated subnet, using the production server and storage technology, the contractor was required to do the following:

- scan versions of the 1996 Census questionnaire,
- create multiple copies of the images,
- populate the image base to a level representing roughly half the eventual size it would attain in production, and
- using two workstations, retrieve volumes of images representative of anticipated production volumes in a specified time frame. This was staged, starting with one questionnaire, then 10, then 100, etc.

Under this scenario, the contractor's solution was required to deliver the first page of a questionnaire within 5 seconds and subsequent pages of the same questionnaire in less than 1 second.

Of equal importance was the integration of the contractor's offering with internally developed production systems. In order to meet requirements, the contractor had to work with a systems development expert to clearly demonstrate that an image viewer (complete with an image) could be launched successfully from the Census production application.

After the tests, Census representatives were reasonably confident if a performance problem occurred that the vendor's solution would not likely be the cause.

On the Census side of the technology, steps were taken to optimize performance.

- The network was configured, sized and properly tested.
- A large RAID device was acquired to cache frequently requested images.
- Pentiums with sufficient RAM and disk were acquired.
- High performance monitors and video cards were acquired.

The net result was satisfactory delivery of images in production. A high level of availability was also experienced as downtime was minimal.

2.6 Why It Worked

When the move to document imaging began for the Census of Agriculture, stories of imaging disasters abounded, no significant imaging applications existed within Statistics Canada and the people tasked with delivery had no experience whatsoever with the technology. A very expensive and visible failure could have happened for the Canadian Census of Agriculture. A reflection provides some reasons why it succeeded.

Statistics Canada's Informatics Organization: Technical expertise for mid-range computers, networks, workstations and system development were available on a moment's notice to contribute to the project. This was just as critical during production as it was in the activities leading up to it.

Statistics Canada's Technical Infrastructure: For some time, the Agency has strongly emphasized adherence to standards. This set the stage for successful introduction of standards-compliant technology into the environment.

Methodologist: A methodologist assisted team members in analyzing the technology and in producing meaningful statistics.

Technical Liaison: Today's informatics projects often require several different technical disciplines — mid-range computer specialists, network specialists, workstation specialists, applications developers, database specialists, product specialists, etc. The imaging project employed a technical liaison that coordinated activities and represented the Census with industry and technical groups.

Turnkey System: Rather than split the contract into say, Company #1 to generate images, Company #2 to manage images and Company #3 to provide image viewing services, a complete imaging system was purchased. This ensured that all pieces of the system would be interoperable; i.e. images produced by the scanner would be compatible with the viewer. It also prevented communication problems in the event that service was required. For the Census, there was only one contractor to call.

Simplicity: The absence of indexing stations and quality control stations reduced the complexity of the application. In addition, a minimal interface between the Census production application and the FileNet Imaging System reduced integration issues and correspondingly reduced integration problems.

Hands-on experience: Sole reliance on literature and demonstrations would have been dangerous, particularly with scanner functionality. In addition, hands-on evaluation of monitors ensured that adequate quality was acquired; it would have been easy to spend too much or, disastrously, not enough. Wherever possible, Census staff acquired loaner technology and gained hands-on experience using the Census documents and environment.

Focusing on the need: Often, throughout the course of the project, it would have been easier to go with conventional solutions such as indexing stations and QC stations. This would have unnecessarily increased costs, complexity and risks.

Perseverance: Several times, throughout the course of the imaging project, it could have been terminated.

Management support and personal involvement: It is widely known that many projects fail due to a lack of commitment by management. The imaging project did not suffer that consequence; far from it. The assistant director and particularly the project manager publicly demonstrated support, fought the political battles and played a major role in shaping the eventual system.

Relationship with the contractor: This was excellent. Keys to success were availability, accessibility and willingness to compromise. It is important to note that the project represented a foothold in the Canadian Federal Government for the contractor. Success was as important to the contractor as it was to the Census. This undoubtedly had an impact on the relationship.

Many contributors: Many people were involved in shaping the imaging system. Project staff were involved in assessing monitors, video cards, image viewers and scanners, sizing the network and building programs to test it, defining the overall structure, building RFPs for production and pilot tests, and defining how images would be used in production. Others fought the unavoidable political battles.

This broad level of involvement cultivated a greater sense of ownership and undoubtedly contributed to the success.

Budget: Acquisition, implementation and maintenance costs are not trivial. A healthy slush fund allowed for unforeseen adjustments to the implementation.

Paper questionnaires were not an option: Often, when technologies are introduced, the technology they are designed to replace still remains. Even if a compelling reason to change exists, unless people see it, they often continue going about their work using the tools they are most comfortable with. For the Census of Agriculture, the 1996 application represented a complete redevelopment. The manual questionnaire delivery system employed in 1991 no longer existed. Only in exceptional circumstances was a physical document drawn in 1996.

Acceptance: Production staff were highly satisfied with the quality of the images and functionality of the viewer. Annotation, zooming and scrolling were considered to be extremely productive features.

Luck: The project was fortunate to have many people on the team and in the service areas blessed with talent. Two or three key project people were remarkable.

3. Reflections and the Future

The goal of timely delivery of images to the desktop was met. Acceptance and appreciation of the technology by all users and observers was positive, to say the least. Imaging combined with an excellent set of edit and validation tools proved to be a powerful combination for users. The only question staff had was, "How did we process the Census in the past without the images?"

The expected benefits were realized and more. Use of images occurred in many processes and applications where it was neither planned nor anticipated. Some of the unexpected benefits were:

- Productivity improvements far exceeded the expected 10 percent level. In many clerical operations, the image viewer automatically zoomed-in on the questionnaire's problem module and the correction was made on-line. This proved to be several times faster than leafing through a paper questionnaire and printouts. In the manual system, correction also included writing the correction on the printout, which was then passed to someone else to make the actual changes on-line.
- During Census collection, it is possible to receive more than one questionnaire for a given farming operation. Procedures are carried out to identify and remove this duplication. Access to the images of the potential duplicates using split screens resulted in significant savings in time.
- An undercoverage problem during collection resulted in a need to identify farms which were missed. To identify them, matches between the 1996 base, tax, surveys and previous Census lists was carried out. Access to the images allowed editors to make a last quick check if the farm was on the base. Field follow-up costs, response burden and the probability of creating duplicate questionnaires were all reduced. Images provided information (e.g. land descriptions etc.) that were not on captured the database.
- Hogs grown under contract for others is a rapidly growing service in Canadian agriculture. Many farmers growing hogs under contract failed to report the hogs on their questionnaires in spite of instructions to do so. The hogs did not belong to them, so they did not include them.

We had the farm but not the hogs. Matches between the 1996 Census and the semiannual hog survey identified farms reporting large numbers of hogs on the survey but not on the Census. Were the hogs missing due to processing or respondent error? A quick check of the images answered the question. Follow-up procedures were then carried out with the respondents to add the missing hogs.

- The automated imputation procedures always seemed to leave a few situations for each module of questions where a donor could not be found. In many cases, the cause was respondent or processing errors. With images, subject matter experts quickly corrected these situations. In previous Censuses, "uninformed quick fixes" were made just to move the questionnaires to the next imputation stage. The time to find and pull the physical documents would have seriously delayed the imputation process.

Did imaging pay for itself in one cycle? No, and nor should it have been expected to. How do you put an economic value on quality of data, job satisfaction, or the cost of a delay in the release date?

Will imaging pay for itself in the short term (5 years)? Without question yes. The equipment continues to be used by the Census for testing, and for many other projects on testing and production. The equipment will also, along with upgrades and additions, play a role in the 2001 Census. Census of Agriculture's positive experience has encouraged others to test and implement not only imaging but ICR applications. Areas now involved in the technology include:

Whole Farm Data Project,
Census of Population,
Tax Data Division,
Industrial Organization and Finance Division,
Operations and Integration Division,
Administrative Support Services Division (Records Management), and
Project to Improvement Provincial Economic Statistics (PIPES).

In some cases the Census of Agriculture has taken on work for other projects. This additional work helps pay for the system, plus share the maintenance and licensing costs, supports related research, expands/maintains experience of the staff, and adds more features and capabilities to the system.

The decision to implement imaging was unquestionably correct. The level of risk taken in implementing was acceptable but not insignificant. Any additional challenges would have put the entire project at risk.

Experience has proven that the risk associated with implementing imaging was well worth taking. In the end, it played a vital role in aiding the project to meet its planned release dates.

A number of events pulled the project off schedule. The undercoverage problems alone had put the data validation task several weeks behind schedule. In addition problems with the generalized imputation system forced the development of a major fix and re-imputation of all the financial variables for two of the largest provinces. A total team effort, well above and beyond the call of duty, plus efficiencies generated from access to the images allowed the release of the data base to take place on schedule.

Where does technology take the 2001 Census of Agriculture? Can ICR handle 16-page legal-sized forms completed by 300,000 aging farmers? Once again there are different opinions and camps. Some

project members who were in the conservative camp in 1996 are in the progressive camp now, and vice versa. A number of other things have changed. We know imaging works and works well. We have more experience, confidence and we have more time.

ICR research for the Census of Agriculture started almost one year ago, four years before the 2001 Census. Early results are encouraging, but it is far too early for any predictions let alone decisions. An assessment of respondent reaction to ICR friendly questionnaires will begin in 1998.

What is the fall back position if ICR does not prove to be cost effective or appropriate for the project? Imaging with more parallel processing is a conservative option. Mark Character Recognition (MCR) combined with key-from-image would be another significant leap forward.

Document imaging is considered by some to be a transient technology as the move to the paperless office continues. A question to ponder is, "When will electronic reporting replace paper questionnaires for the Census of Agriculture?"

Related Documents

Bradshaw, C. and Duggan, J. (1997), *A Review of Document Imaging for the 1996 Census of Agriculture*.

Duggan, J. (1996), "Electronic Imaging in Support of the 1996 Census of Agriculture," presented at the International Conference on Computer-Assisted Survey Information Collection.

Green, I. (1995), "Questionnaire Imaging Pilot," internal report, Statistics Canada.