Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

2006 CENSUS DATA CAPTURE REDESIGN

Jean-René Boudreau and Timothy Withum[1]

ABSTRACT

Data capture is one of the essential steps in all surveys. If it is not done with great care, it could generate non sampling errors which may invalidate analyses. The Canadian Census of Population has used the facilities and personnel of the Canada Customs and Revenue Agency (CCRA) since 1981. This collaboration has been beneficial to both agencies. Data capture, that is the transfer of the information from paper into an electronic medium, was carried out by keying, also referred to as "direct data entry". Since this operation is very repetitive, it is subject to random errors. As a result, statistical agencies are looking at technologies where the transfer of information of most fields would be done by a machine. In the last few years, Census managers have redesigned the collection and processing activities of the Canadian census to satisfy a strong need for integration of multiple collection methods: mailback questionnaires, Internet, computer assisted telephone interviews, etc. One of the activities that will change considerably is data capture. Through a government competitive process, Lockheed Martin (LM) was awarded the contract for the data processing of the 2006 Census, including a data capture system using optical technologies. Their system will scan, recognise, process and save most of the information largely without human intervention. It is also important that LM maintains the quality obtained by CCRA in previous censuses. In this paper, we discuss the challenge of moving from a world of keystrokes to a world of character recognition.

KEYWORDS: Direct Data Entry; Intelligent Character Recognition.

1. INTRODUCTION

To be able to adequately explain to Canadians how their society is evolving, which is one of the main objectives of the census, the data capture activity, and more generally all census operations, should put as much energy as possible in reducing the level of non-sampling errors. The data capture operation is crucial here because its purpose is to transfer information from the paper medium to an electronic one. We know that changing the support medium always introduces noise into the information. If we are not careful, the information content may be irretrievably lost. Significant resources are therefore allocated to this activity.

When a data capture methodology is changed, the challenge is to maintain the quality of the data while almost all factors that have impact on it will change. Technologies, allocated budgets, quality assurance programs, the questionnaire design, etc., are all major factors on the quality of data capture and yet we want the output from the new methodology to be of equivalent or better quality than from the previous one. Of course, the choice of a better methodology for capturing the data should result in the elimination or a significant reduction of old, chronic problems driven by the previous methodology. In the case where a difficulty remains in both methodologies, the new methodology should provide better quality, or at least, potentially better quality in forthcoming versions of that methodology. And, since nothing is perfect, as the new methodology will undoubtedly create some new problems, it is expected that a better methodology would not create more problems than it would solve.

This paper discusses the challenge of changing from the direct data entry methodology used for the 2001 Canadian Census of Population to the optical character recognition methodology proposed by Lockheed Martin to capture the 2006 Canadian census. To discuss the matter intelligently, we propose the following outline. We roughly explain each methodology by emphasising their strengths. We then focus the discussion on different problems or difficulties that each methodology faces and describe how they are handled. Then finally, we examine the impact on the data quality. After this exercise, the reader should be in a good position to assess whether or not the challenge has been well worked out.

[1] Jean-René Boudreau, Statistics Canada, Ottawa, Canada, K1A 0T6. Timothy Withum, Lockheed Martin, Beltsville, Maryland, USA, 20705

2. DIRECT DATA ENTRY

Statistics Canada has contracted out the data capture operation to the Canada Customs and Revenue Agency (CCRA) in every census from 1981 to 2001 inclusive. It was a decision that benefited both agencies. For CCRA, every five years the census was bringing in an additional workload in a timeframe, June – November, that was not critical for CCRA. The benefits for Statistics Canada were the use of CCRA's infrastructure (space, equipment, IT resources), and the use of CCRA's experienced employees. This cooperation resulted in major savings to the Canadian taxpayer.

A few numbers may help the reader grasp the context of this operation. The reader may consult Boudreau and Liu. (2002) for more information. The program for the 2001 census started June 3rd 2001 and was completed by October 25th. Eight CCRA processing centres across Canada ran the program. Around 1,700 experienced keying operators participated in the program. They processed over 13.2 million questionnaires, which held approximately 600 million fields to capture. To process this quantity of information, they keyed over 3.8 billion keystrokes with an average of 104 keystrokes a minute. The important question is: How good were they? Compared to standards, they were excellent. The error rate at the question level was 0.74%. At the field level, the error rate was as low as 0.23%. We should digress a little here. How does one calculate an error rate? It is the proportion of units (i.e. fields, questions, questionnaires) that are incorrect over the total number of units that were processed or should have been processed. The word "incorrect" means that what we find on the electronic medium after processing is different from what we find originally on the paper medium. Also the system processes a unit only if there is some information; an empty field is not processed in that sense. In addition, when we say in the definition: "or should have been processed", we mean the inclusion of all units containing information that were erroneously missed by the data capture system.

How did CCRA operators capture the information? To answer this question, we have to describe roughly how the questionnaire is designed. A population census questionnaire is a 3-dimensional matrix. The first dimension specifies the pages (both sides) of the questionnaire, which essentially is a set of questions. The second dimension determines the person, in pages where person questions are asked, or the page, in pages when household questions are asked. Each block is labelled by a number we call the "block ID". The third dimension determines the question. It is essentially a set of fields. Each field is labelled by a number we call the "cell ID", so any field inside the questionnaire has a unique set of coordinates: block ID x cell ID. There is no fixed location on the capture screen to put specific information; operators use these coordinates to enter the data. The keying is therefore long strings of characters. A building block of a string is made up of three components: the block ID (if necessary), the cell ID of the field, the content of the field (if necessary), and a field separator. Each non empty page produces a string. For example, if data in a questionnaire looked like this:

67.	68.
01 q	01
02	02 q
03 ENGLISH	03 FRENCH
04 52	04 25
!	!

the string "67.01-03ENGLISH-0452-68.02-03FRENCH-0425-" should be keyed. The system uses the period "." to identify block IDs (the two characters before any period). The dash "-" is the field separator. The two characters that follow a dash and are not block IDs are cell IDs. The characters that follow cell IDs form the content of the field (i.e. the response). If a cell ID is followed by a separator, it means that the respondent answered the check-off circle field identified by the cell ID.

This is by no means an obvious way to capture the information from a questionnaire. We would not recommend this approach if the keying operators were not familiar with it. But CCRA operators are accustomed to it. It is natural for them because the data capture system used is a modification of one of their data capture systems. It is very fast because operators can enter long strings before sending the transaction to the operating system. In fact, operators fill their screens, which consist of several pages, before submitting transactions. Operators are therefore focussing on questionnaires. They do not have to watch their screens while capturing information.

Obviously, the capture system is performing some verifications to ensure that information captured is accurate. There are three levels of verification: the block and cell ID validation, the confirmation process, and the key verification.

A. Block and Cell ID Validation

The data capture system knows everything about block and cell IDs. If an operator enters a block ID that does not belong to the page currently being processed, then the system does not know how to interpret the string. The same thing can be said for cell IDs. When this happens, the system prompts the operator to rekey the entire string (all the page information). When all block and cell IDs are valid, the system analyses the string. The block and cell IDs in a string have to be in increasing order. For example, if cell IDs are not increasing within a block, it may be because the operator forgot to insert a block ID when he/she started a new block. When this analysis fails, it is always a data capture error. The system then prompts the operator to rekey the entire string. It may be said that we are penalising operators too much by recapturing the entire string. This is not the case. It is faster and easier to recapture everything. Otherwise, operators would focus more on their screens, reducing their concentration on the actual keying.

B. Confirmation Process

Every response goes through a series of edits. A response like "80" for the number of weeks worked in a year, or an invalid postal code, are examples of what this verification is intended to find. When a response fails at least one edit, then there is either a capture error or the response is "judged" questionable. In all these cases, the system asks for a confirmation. Later in the process, the system will prompt another operator to rekey the same response. If a difference is detected between the two keyings, the system will again prompt the 2nd operator to rekey the response. It makes the 2nd operator the verifier, and the referee, if needed. In the 2001 census, 630,000 (= 0.11%) fields contained a response that failed at least one edit, but were later confirmed. Also, just over 1.3 million (= 0.22%) fields contained a response that failed at least one edit, but were in fact, captured erroneously. This percentage is of the same order of magnitude as the final error rate at the field level. (Thanks to that verification!)

C. Key Verification

The purpose of the Key Verification (KV) is to control the error rate at the field level for all the fields that passed all edits of the previous verification. This verification is a formal quality control (QC) used to ensure that the error rate stays below a specified threshold. As in any acceptance sampling QC, the questionnaires are put into batches prior to the original keying. A certain number of questionnaires are randomly selected in every batch. These questionnaires are rekeyed completely by other operators (KV operators). Every time there is a discrepancy between two keyings, the KV operators are the referees; the system asks these operators to rekey the information until the system is able to determine who made the mistake. Since KV operators are referees, CCRA management has always allocated their best operators to do this verification. If, for a particular batch, there are too many mistakes made by the original operator, then the system asks the KV operator to rekey the remaining questionnaires. The keying of these questionnaires is the same as if they were already identified in the sample. In 2001, around 9% of the questionnaires were selected in the sample. A little less than 11% of batches were rejected.

Just over 25% of the total number of keystrokes were keyed in these verifications.

3. INTELLIGENT CHARACTER RECOGNITION

In this section we describe the Intelligent Character Recognition (ICR) methodology. There is one difference, however, that should be mentioned. The description in this section is a proposed solution for the data capture of the 2006 Canadian census, not a description of a methodology already used. The level of precision or accuracy of the description may not be comparable. There are even some procedures that are still under review. Sometimes, we do refer to the Lockheed Martin experience with the UK and US censuses.

The objective of this methodology is to perform the easy and repetitive processing by a machine. The rationale is that data capture is an activity that is composed of millions of little, mechanical tasks. The Lockheed Martin proposed solution is a design built to process the majority of these tasks very reliably, and to send the less reliable ones to a process where more resources are allocated (i.e. human intervention). The methodology can be split into four parts: scanning of questionnaires and image processing, mark and character recognition, key from image and paper, and the quality assurance program.

A. Scanning and Image Processing

The first process scans the questionnaires to obtain an image of all the sheets. In some cases, since questionnaires are booklets, they have to be guillotined first. Once the paper is scanned, the image processing system aligns the images with a template so all the fields are in a consistent location for subsequent processing. As part of that process, an assessment is performed to ensure that the image is of sufficient quality to allow highly accurate and efficient data capture. In cases where the assessment notes a quality failure, the error can be corrected by an operator or the paper can be pulled and rescanned or keyed from the paper form.

When the image of a sheet is satisfactory, the system runs a processing series to "wash away" unnecessary marks and colours outside response areas. This is done in two stages. The first stage is really through form design. By choosing the proper colours (ones that "drop out"), the scanner itself washes away the colour. The second stage operates on the resulting black and white image. This stage overlays the template image on top of the scanned image and essentially erases the template from the original. What remains on the image are marks, numbers and words in or outside response areas that are not part of the template.

B. Mark and Character Recognition

There are two engines used to recognise responses: the optical mark recognition (OMR) and the intelligent character recognition (ICR). OMR is used to register if a check-off circle was checked, and ICR is used to decipher a write-in answer. The recognition algorithms work with the number of pixels that are set (= 1) and their location in the response area. OMR registers only the number of pixels that are set. Multiple levels of mark sense detection are employed in order to achieve accurate results. The first level is a simple pixel thresholding algorithm. The engine converts the number of pixels that are set into a value with a specific range. This value is called the "confidence value". If the value is higher than a specified threshold, then the engine records a mark. If the value is less than another specified threshold, then the system does not record a mark (the field is viewed as empty or blank). If the value is between these two thresholds, then the engine requires some help from human intervention. For the ICR engine, the same kind of process is used although it is much more complex. First, ICR has to segment the response in characters. Questionnaire design is a big help. Every write-in response area is split into boxes of equal dimensions to discourage the use of cursive writing. Respondents are asked to use capital letters. After the characters have been separated, each one has a confidence value. A series of tests are conducted at the field level to increase or lower the character confidence values. For example, in both English and French, some combinations of letters may increase their confidence values (e.g. a "q" followed by a "u" in the string). Three "i"s in a row ("iii") does not frequently occur, therefore this series will decrease their confidence values. Also, if the number of responses is limited, such as province or country of birth, the system may look up responses in a reference table, also called a dictionary. This table contains series of characters that look like possible responses. When a match is found, then it greatly increases the character confidence values. It may be worth noting that ICR will often produce second choices for each character. For example, the written letter "C" may be recognised as a "C" with the second choice at a lower confidence as a "O".) The dictionary processing will take those second choices into account during the lookup. After these analyses are done, if the confidence value of the field (the lowest confidence value amongst all characters in the field) is higher than a specific threshold, then ICR registers the string of characters as the response. If there are not enough pixels that are set in all response boxes, then ICR does not record an answer (the field is viewed as empty or blank). In other cases, ICR requires some help. As in the case of OMR, it needs human intervention. Roughly, this is how the recognition engines work. From past experience, 0.1% of marks and 20% of write-in responses need human intervention; although the last percentage greatly depends on the type of fields.

C. Key from Image and Key from Paper

When OMR or ICR needs human intervention, the field has to be keyed by an operator. This operation is called Key From Image (KFI), sometimes referred to as "data repair". This operation may be viewed as a "queue service line", where each field goes into a queue to get "serviced" (keyed). The KFI screen is very particular. Its purpose is to help the operators remain focused on the screen. The screen is divided into three identical horizontal sections. The section in the middle contains information on the current field to be processed. The section above contains the information of the last field being processed and the lower one contains the information of the field that will be processed next. Each time an operator processes a field, the sections are updated. This screen allows operators to reconsider the field that was just processed and/or to have a quick look at what is next. This method has proven to be very effective. In each section, there is an image snippet around the field and a capture area. The operator is supposed to key what he/she sees in the field shown in the image snippet in the capture area. Again, the fields that go into this

process were not clear enough to be recognised, so strange responses are expected. This means well defined capture rules are necessary (e.g. What does one key when two numbers are present in a field?).

There will be some questionnaires that cannot be scanned because they are somehow damaged. Other questionnaires will not give clear images because, for example, their responses bled through the paper severely, interfering with responses on the other side of the sheet. When this is the case, all the sheets of these questionnaires will have to be keyed by some kind of direct data entry. This operation is called Key From Paper (KFP). To enter the data, the system will provide a modified version of the Internet screen that will be used by Internet respondents. Operators will be asked to answer the questions on the Internet screens by entering the information on the questionnaires.

D. Quality Assurance Program

The objective of the quality assurance program is to ensure the same accuracy as the DDE data capture system of the 2001 Census of Population. The confidence thresholds described above, the point that divides an automatic recognition from a manual resolution, is of crucial importance. If that threshold value is too high then, although the recognition will be very accurate, KFI will be solicited too often. Timeliness and resourcing will be problematic. On the other hand, if the threshold value is too low, then quality may suffer. The confidence thresholds need to be properly set to maintain an equilibrium between quality and resources. The available technology permits raising these threshold values high enough to balance both objectives. Since the capturing of information will mostly be done by machine, the machine must first be well tuned. Also, the quality of the KFI and the KFP operations has to be monitored to guarantee the data quality to the prescribed level.

How does Quality Assurance monitor the OMR and ICR engines? It challenges the system by asking for a second pass on fields that were recognised automatically. To do this, a sample of fields is selected. The system reduces the confidence values of all the fields in the sample that were recognised automatically. These fields then go into the KFI process where they are all recaptured. The initial and the second keyings go into an arbitration algorithm to determine if the initial keying is right or wrong. Accuracy rates are calculated and actions are taken if these rates are too low.

The second thing to monitor is the quality of the KFI operators. To accomplish this task, the system randomly selects fields from batches again. Here a batch is the unit of processing, a set of questionnaires. If a selected field has a low confidence value, it is sent back to KFI to be rekeyed by another KFI operator. Then the system takes what the engine recognised first (even though their confidence values are low), the initial keying, and the second keying from KFI, and sends these three pieces of information into an arbitration algorithm to determine if the first keying in KFI is right or wrong. If the accuracy rate observed in the sample of a batch is too low, then all the fields with low confidence values are rekeyed in KFI. A fine feature of this process is that KFI operators do not know if they are capturing, verifying or arbitrating the information. The quality assurance of the KFP operation is similar except that the sampling unit is an entire questionnaire rather than an individual field. (While the field-level sampling is preferred to get a more efficient sample, this is not cost efficient in KFP.) After a series of batches of questionnaires are keyed by a KFP operator, a sample of questionnaires is selected to be rekeyed by a second operator. When this rekey is done, the system sends the initial and the second keyings of the selected questionnaires to an arbitration algorithm. If the accuracy rate of the first KFP keying is too low, then the batches are completely rekeyed by a KFP arbitrator. This special KFP operator rekeys all fields in all questionnaires in the batches and arbitrates when necessary. If his/her keying of a given field is different from the original KFP keying, the system prompts the operator to rekey the field in question. The actual mechanism of evaluating batches in the KFI and the KFP operations is a bit more complex than explained, but the simplification does not change the overall scheme.

4. ANALYSIS OF THE CHANGE IN THE DATA CAPTURE METHODOLOGIES

As we said earlier, the impact on data quality will be established by discussing how each methodology tackles different problems or difficulties that might arise. We can arrange this section in three parts. The first part deals with difficulties in the DDE methodology that should be eliminated by the design of the ICR. The second part discusses problems which both methodologies confront. In that part, we focus on the simplicity of possible solutions. The last part discusses the difficulties that ICR has to deal with, but are not really problems for DDE.

4.1. DDE difficulties eliminated by ICR

A. Errors of coordinates (block ID x cell ID)

The first error that ICR eliminates is the keying of a wrong cell ID that is undetected by the system. If we use the same example as before, the operator should have keyed: "67.01-03ENGLISH-0452-68.02-03FRENCH-0425-", but keyed: "67.02-03ENGLISH-0452-68.02-03FRENCH-0425-" instead. The cell ID "02" in the block ID "67" is a valid cell ID. This situation generates two data capture errors. The first error is a missed field (the right cell ID is not keyed), the second is an extra field (the wrong cell ID is keyed). From the QC sample, we observed that 17% of all the errors that could not be detected by the system were, in fact, extra fields. We believe that a significant amount of extra fields were produced in this way in 2001. More generally, with human processing, it is possible for a person to create data that is not actually on the questionnaire. They may have been looking at the wrong field when they were keying, etc. In the case of ICR, it is almost impossible to create data from nothing or completely delete data that was present. At worst, the field is sent to KFI for resolution.

B. Special characters that cannot be keyed

ICR can capture the periods "." and the dashes "-". DDE cannot capture these characters because they are used by the system to identify block IDs and to separate fields. This is important because it reduces the number of instructions for KFI and KFP operators. The motto: "Key what you see" includes all these cases in ICR. In the DDE methodology however, the instructions are to key a space " " every time a period has to be keyed. This is especially confusing in fields with a lot of abbreviations (e.g. U.S.A. Should we key "U S A" or "USA"? Or should we key two spaces if a period is followed by a space?). These types of situations are not generating many errors, but operators do have difficulty taking "Key what you see" seriously. The major constraint with the inability to capture a dash is: "What do we do with a negative number?". Since ICR can key these characters, it reduces the number of special cases to process, which is very important to any data capture system.

C. Truncation of responses

Truncation is another difficulty that will be resolved by the ICR methodology. The following is very specific to the DDE capture system developed by CCRA. The way the information was exported to Statistics Canada (one physical record per document) limited the number of possible bytes. Each alphanumeric field had an allocated number of bytes determined by negotiations. Over 50 million alphanumeric fields were keyed. From these, 0.19% were truncated. However, the truncation was not distributed equally among questions and languages. For example, 0.99% of the English / 1.68% of the French "Main activity" fields were truncated. The truncation issue is important because some of these fields may contain important words at the end of the responses. This is different than not having enough boxes to enter the information in the ICR methodology. If that happens, then a respondent may print two or more letters in the boxes or will continue writing the answer near the boxes. These cases should receive low confidence values and will be sent to KFI operators. These operators should be able to enter all the responses' content.

D. Referral activities

There will be no referral grooming activities in the ICR world, (activities that modify the data before the actual keying). In the DDE world, the referral clerks are supposed to "correct" the data before keying because its current format is not acceptable. If the requested "monthly rent" is given for the whole year, the operator is supposed to ask the referral clerk to make the appropriate conversion. The original answer is crossed-out and the converted amount is written near the response area. Referral clerks write on questionnaires to avoid redoing these grooming exercises when these fields are rekeyed (if needed) in one of the verification processes. It makes the verification and the arbitration process very, and probably too dependent, on the original keying. With ICR, since operators work with images, independent processes occur if the same fields are rekeyed. Also, referral clerks do a lot more than conversions. They are supposed to apply all the grooming rules in the operator manuals, for example: What should be captured when two sources of income are reported for an income component? They even have the final say if words are not legible. Overall in 2001, 0.13% (= 425,000) of all questions needed referral clerk interventions. This is still a sizable number of corrections. Although the ICR methodology solution for these keying problems are not yet finalised, any solutions that are close to the "Key what you see" motto should be preferred: "Enter the data and let the system do the math, or more generally, do the processing". In the example of two sources of income, keying the two numbers and separating them by the plus sign "+" does the trick. The referral clerk is replaced by the insertion of the two entries and by the introduction of some code source to add the two numbers together in a post data capture processing.

4.2. Common difficulties

A. Illegible answers

Keying procedures for illegible answers are somewhat difficult because they contain elements that are contradictory. For example, in the 2001 census, the procedure for illegible answers was: *"The referral clerk should try to determine what the response is and to cross-out any part that is illegible. If the answer is completely illegible, cross-out the entire field"*. But further in the manual the procedure also specified: *"Do not attempt to interpret what the response should be if the respondent provided part of a response"*. The words "determine" and "interpret" have close meanings in this context. We anticipate that in 2006 all these cases will be sent to KFI for a rekey. If these fields are recaptured by other KFI operators, independent evaluations will be made. That may be better than only one opinion.

B. Conversions

Conversions occur when reported amounts are not in the requested unit of measurement. Yearly rent instead of monthly rent, hourly wages instead of yearly wages, etc. are some examples. In the DDE methodology, the referral clerk has to do these conversions. In 2001, there were 180,000 (= 0.17% of all numeric fields) numeric conversions identified. In 2006, they can only be identified if there are some measurement units in or close to the response areas (e.g. per hour, per week, etc.). Without these units of measurement, conversions cannot be done. Also conversions may be identified if the responses fail some range edits. The observed error rate for conversion was 3.97% in 2001. If ICR determines that a conversion is necessary, the KFI operator may not be able to do the conversion because only snippets of questions are displayed on KFI screens. The KFI operator can only key the reported amount and indicate the conversion required. Again, the ICR methodology solution for conversions is still under investigation but, as we have mentioned in the previous subsection, any solutions that would ask operators to key the data and let the system do conversions is much better than asking operators to clean up the data.

C. Substitution Errors

In the ICR world, what we hear most often is "substitution errors"; the ICR engine incorrectly recognises a character. This happens in the DDE world too but at a lower degree. In 2001, the estimated error rate caused by wrong keying for all write-in fields was a bit below 0.07%, which represents an accuracy rate of 99.03%. These errors depend directly on the reliability of the keying operators in the DDE methodology. In the ICR methodology, the reliability of the ICR engine has to be added to the reliability of KFI and KFP operators. The US census experience showed that, for the ICR engine and the KFI operators, the accuracy reached 99.5% and 98% respectively. That gave an overall accuracy of 99.4%. We should be prudent with these numbers. They were calculated with different error definitions for alphanumeric write-in fields that benefited CCRA operators. It is not possible to adjust them to make a fair comparison.

D. Respondent's cancellations

Should we capture fields that are crossed-out by respondents? No, we should not. In 2001, there were 2.5 million questions that contained at least one crossed-out fields (= 0.75% of the total number of questions). Although the procedure for field cancellation was clear in the DDE methodology, it brought a lot of confusion because it is not always easy to determine if a specific field is cancelled. ICR has the same problem. If the cancellation occurs in a write-in field, then the confidence value is expected to be low and the KFI operator that processes the field can decide if it is cancelled. For check-off fields, it may not be as easy. The system extracts more complex features from each mark to determine whether or not they are realistic enough to be considered the respondent's intended entry. Each of these factors are evaluated in order to classify each mark as real or spurious. As an even further method of evaluation, multiple marks on a sheet are collectively evaluated in order to determine the respondent trend for providing a mark. The accuracy for OMR in the US census experience was 99.98%, which is very similar to the DDE 2001 experience.

E. Visual Signs

Sometimes respondents write visual signs instead of answers. Since population census questionnaires are designed in a matrix form, a respondent may draw an arrow or use ditto marks from his/her answers to indicate the same answers to other persons on the same page. Or, he/she may answer a question with a question mark "?" in case the answer is unknown. In 2001, 0.20% of all questions showed these types of signs in their response areas. In the case of an arrow or ditto marks, the keying operator was instructed to key what the sign meant. That procedure did not produce an accurate keying for these fields. The error rate for these questions was 11.38%. The keying procedures

were not uniform; some were different for different questions. The occurrence of these signs depends on the questionnaire design. The matrix form makes it easy because each response area of all persons are on the same line or row. The introduction of segmented response areas to help the ICR character segmentation may reduce their occurrences. Since the census is a self-enumeration survey, too detailed questions (like income components to the dollar) accentuate the occurrence of a question mark "?" in a numeric field. For 2006, if one of these signs is present in a field, we expect that its confidence value will be low and the KFI operator will be able to make the appropriate action. We believe there is possibility of reducing the error rate.

F. The Human Factor

The use of ICR methodology also reduces the need to hire and train a large temporary workforce. The DDE methodology requires a highly trained staff and CCRA may not be in a position in 2006 to generate a workforce of equal size and quality as in previous censuses. Although there is still a challenge in staffing and training a data capture workforce, the ICR methodology, because of the reduction of its workforce size, reduces that problem significantly. A fine feature of ICR/OMR engines is that they are not affected by stress, long hours, etc. Also, once trained, multiple copies of the engines can be replicated across the system (with no learning curve) in order to increase processing throughput. (Additional people will often require training.) Because it is automated, the engines can use information from the entire questionnaire in order to improve accuracy. For instance, the algorithms can use all the surnames on a questionnaire – which may be the same – to help determine the proper surname to be captured. Operators can do this too, but it takes much longer.

4.3. New problems for ICR

A. Information outside response areas

There were 1.2 million questions (= 0.20%) that had a part of the response, or the entire response, outside the responses areas. Even if the error rate for these questions was high, at 4.9%, only 14% of these were missed questions. The DDE is a good methodology to at least identify them. There were even procedures to follow if the respondent answered a question by indicating examples in the question area. How can ICR identify answers outside response areas? When an image of a sheet is made, several processes wash away all information contained in the template, the form drop out as described above. After these processes, if the sheet is well positioned in the scanner, the resulting image will show only writing by respondents. An analysis of the number of pixels that are set outside the response areas is done. The system can decide that a specific sheet has too many of these pixels. Then the action could be to rescan the sheet or send all the questionnaire sheets to KFP for a better capture method.

B. Questionnaire Integrity

The difficulty here is to ensure that the guillotined sheets of a questionnaire will remain together. If all questionnaire sheets had a unique bar code identifier per questionnaire, then there would be no problem. Sheets could be processed properly even if they were shuffled. However, this condition is often too expensive. Alternative methods of document integrity are available that are nearly as accurate but much less expensive. Explaining these methods is however beyond the scope of this paper. It is still under investigation.

C. French Accent Recognition

The French accent recognition has never been a difficulty for the DDE methodology. The keying operators were instructed to key only letters and not accents if any. For the ICR methodology, it may not be as easy. French accents may perturb the character segmentation algorithm because the "universe" of characters to recognise is larger.

5. IMPACT ON DATA QUALITY

The quality of the data must be assessed after every major activity in a survey. The data capture of the 2006 Census will not be an exception to that rule. A study will be conducted after production to evaluate error rates and to determine the reasons of these errors. However, we are now asked to assess the impact of the change of methodology on data quality, which is not the same thing. Data quality is usually measured in terms of variance and bias components. In this context, variance is a function of the reliability of the methodologies in reproducing the same results from similar conditions. A simple example is a measure of dispersion of the keying when we ask DDE operators to key the digit "5". For the bias component, another measure is required. This time, we ask about the accuracy of the

methodologies confronted with a given condition. An example of a bias in the DDE world would be: all referral clerks in one processing site systematically make the same error interpreting a keying rule. In the ICR world, an example would be: the responses outside response areas are very different from those inside response areas and they are all missed by ICR. We discuss here the plausibility of reducing or enhancing data quality by changing the methodology. We separate the analysis in two parts. The first part deals with straightforward data capture, where there is no need for any correction. The other part discusses when correction is required.

A. Straightforward Capture

When DDE keying operators make mistakes, which component do the errors fall into? Do these errors produce mostly variation or bias? From the position of the hand over the numeric pad and also from the positions of the numeric keys, can we observe some pattern when operators enter digits? It is hard to tell. The errors seem to be random. In 2001, 66% of all the errors (= 1.6 million) made by DDE operators were missing or keying extra fields. These errors are explained by the reliability of the operators entering block and cell IDs or by inattentive operators. These fields do not represent difficult cases. Mostly these errors go into the variance component. However, these errors can create great distortions for questions with very small field occurrences, like the "Aboriginal Ancestry" question. From the QC sample of the Key Verification in 2001, we observed that the category "Inuit" was keyed 907 times in the sample by the DDE operators in the original keying. The QC revealed that 99 of these cases were really in the category "Not aboriginal", which means that 10% of the number of Inuits were not really Inuits. The majority of these errors were corrected by the QC or were found later in census processes when compared with other variables, like immigration, language, etc. For these questions, we can detect some bias. Thankfully, all these situations will disappear with the ICR methodology.

What is left is essentially substitution errors in the response's content. In 2001, 30% of the errors (= 725,000 errors) made by DDE operators were wrong keying. We believe they are all included in the variation component. The ICR engine has also the same kind of patterns. For example, the digits "8" and "6" are quite similar. The distribution of the digits in error are however far from uniform. But, if bias is somehow detected during production, a change to the processing algorithms may be possible and would almost instantly correct the problem. Also, if a bias is detected after production, it would be known that the ICR would have consistently processed the situation the same way every time. There would be no variation from one operator to the next. Such a situation may be easily correctable.

In the situation of a straightforward capture, which includes 98.17% of all questions to be processed, we conclude that :

1. The DDE methodology leads to mostly variation errors. Two third of them are missed fields and extra fields. We do not believe that the fields in question are more difficult to capture. The ICR methodology will eliminate most if not all of them. This is the main strength of the ICR methodology.

2. The ICR methodology produces mostly substitution errors. These errors will be concentrated on fields that are more difficult to decipher. Since distributions of the digits in error are not uniform, bias is present, but may be "workable". However, any fair assessment cannot be made from the available data because of too many uncontrolled aspects are present (including different accuracy calculations, the keying rules to be applied, the operator's working station), it is too early to conclude that one method is better than the other. This will be investigated further during the dress rehearsal activity when more information is available.

3. There is potential of a better data quality if the keying procedures of the KFI operators are as close as possible to the "Key what you see" motto.

B. Corrections are needed before Capture

There were 425,000 questions needing corrections prior to capture in 2001. Of these cases, around 85,000 questions (= 20.12%) were not captured adequately because corrections were not done properly. If we assume there was only one field per question that needed correction, then wrong corrections would count for 6% of all fields in error; it is not marginal. We do not have data to support what kind of responses trigger corrections but we cannot assume they are similar to the ones that do not. For example, if a person reports his/her wages in $-amount per hour then, although not always true, low yearly wages are expected. If these corrections are not done, bias will therefore be present, surely marginal at the Canada level but may be important for small geographic areas. The ICR methodol-

ogy has not yet finalised its solution to process these cases. If some functionalities of KFI were developed in the line of this paper: special function keys to process conversions, algebraic expression resolutions, etc., then data quality for these questions would be improved significantly.

6. CONCLUSION

Census management has reengineered some collection and processing activities to satisfy a need for centralisation and integration of operations. In the heart of that restructuring lies the data capture operation. With this new way of doing the Canadian census, it is no longer possible to continue with the Direct Data Entry (DDE) methodology designed by the Canada Customs and Revenue Agency to capture the data. A new methodology based on Intelligent Character Recognition (ICR) is mandatory. The data capture system proposed by Lockheed Martin is the one Census management selected through a government competitive process.

We have described both methodologies, their strengths and their weaknesses. Some inherent problems of the DDE methodology will disappear, like missing fields, extra fields, errors with coordinate ID keyings. With ICR, a few new problems arise like questionnaire integrity and information outside response areas. We have shown an imbalance between the elimination of old problems and the introduction of new ones. Given that both methodologies give similar accuracy, the difference is a shift in the kinds and reasons of errors. There will be fewer errors from “easy” fields and possibly more errors from “difficult” fields. We will be in a position to evaluate that shift in the Dress Rehearsal test of 2004, and then take action, if necessary, to minimise its impact for the 2006 census by, for instance, fine-tuning the engines with real data or by adding functionalities to the KFI operation to deal with more complex situations.

At this point, we feel confident that the data quality resulting from data capture will not be compromised. From Statistics Canada experience with DDE data capture and Lockheed Martin’s experience with ICR with the US and UK censuses, the data capture system that will emerge to process the 2006 Canadian census will make the risk of a reduced data quality marginal. On the positive side, we anticipate there is a greater possibility that data quality will be enhanced.

REFERENCES

Boudreau J.R., and L. Liu (2002), “The 2001 Direct Data Entry Evaluation Study”, unpublished report, Ottawa, Canada: Statistics Canada.