

Article

Symposium 2008:
Data Collection: Challenges, Achievements and New Directions

Generalized Quality Control for Optical Data Capture at Statistics Canada

by Lampson Nguyen, Keith Davis, Cameron Oddy,
Hélène St-Jean and Lillian Melki

2009



Generalized Quality Control for Optical Data Capture at Statistics Canada

Lampson Nguyen, Keith Davis, Cameron Oddy, H el ene St-Jean and Lillian Melki¹

Abstract

Statistics Canada has embarked on a program of increasing and improving the usage of imaging technology for paper survey questionnaires. The goal is to make the process an efficient, reliable and cost effective method of capturing survey data. The objective is to continue using Optical Character Recognition (OCR) to capture the data from questionnaires, documents and faxes received whilst improving the process integration and Quality Assurance/Quality Control (QC) of the data capture process. These improvements are discussed in this paper.

Key Words: Data processing; Improving process quality; Quality control; Optical character recognition.

1. Introduction

Following the successful introduction of generalized Optical Character Recognition (OCR) data capture, it became evident that quality control (QC) procedures played an important role in ensuring the success of data capture operations. Although they provided valuable information, the QC procedures (developed outside of the data capture system), were not an integrated part of the collection stream which made it more difficult to maintain, and therefore affected our ability to implement a comprehensive QC program.

Recently, Statistics Canada's head office collection area implemented changes to the QC systems for its data capture operations. The challenge was to customize commercially purchased software, OCR for AnyDoc, including its existing Auditor feature, to develop a completely integrated generalized QC program to fulfill our requirements.

The purpose of this paper is to describe the enhancements made to the data capture and the generic QC approach, and to show the benefits and impacts when using OCR. These enhancements focused on the statistical methods and procedures while ensuring a higher level of confidence, quality, cost-savings and timeliness of the data capture process.

2. Quality objectives

The principal quality objectives for this project were: firstly, to continue to measure, control and improve the quality of data capture operations on a continuous basis; secondly, to add efficiencies to the quality control process in order to see more timely results and minimize system maintenance, and finally to improve the development environment by using pre-existing software modules and reduce code customization. These objectives were achieved by:

- a) Improving the efficiency and data representation of the QC sampling strategy through the use of the systems' sampling functions.
- b) Replacing the many manual procedures within the QC process with automated processes.
- c) Modifying the QC verification step so it is integrated with the production capture process by using built-in software functionality.

The QC design and procedures were adapted to the AnyDoc software while making use of the functions available in the Auditor module.

¹ Lampson Nguyen, Keith Davis, Cameron Oddy, H el ene St-Jean, Lillian Melki, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6

3. Present system

The process for implementing a new paper survey to be data captured and verified using OCR imaging technology requires the following preparatory steps: questionnaire design, questionnaire template development, and template testing. Once these steps are completed the data capture and QC process can commence. The following sections describe these steps in more detail as they relate to the present system.

3.1 Questionnaire design

The key to ensuring successful data quality is to ensure the paper questionnaire is designed in the best way to be used with OCR technology. There are many simple modifications that can be made to a questionnaire which will significantly improve the accuracy of data recognition. This includes using barcodes for document identification, modifying data input areas to minimize the need for handprint recognition (e.g. use checkboxes for data input), utilizing colour for data separation, and printing special markers on a questionnaire to help with image alignment. Statistics Canada has developed guidelines on designing a questionnaire to ensure optimum OCR accuracy and this documentation is provided to survey clients (along with feedback based on their draft questionnaire designs) at this stage.

3.2 Questionnaire template development

Upon reception of the final printed questionnaire and the final capture specifications, the questionnaire template development will begin. This template is required because the questionnaire being processed is considered a structured document (data placement on the questionnaire is static), and therefore the exact location on the questionnaire where the system expects to retrieve data is known. This template is based on each page using digitized images of the scanned questionnaire, and requires identifying all data to be retrieved and key information on each data field. For example, the placement of the field on the page, the type of data (e.g. numeric), the order in which to capture and output the data, the field edits (e.g. check that the value falls within a certain data range), and the confidence levels that the system will base its confidence on when attempting to interpret data. During template development, all required QC settings are added such as, QC parameters, sampling and verification functionality, and QC statistics output and decision results. Once development is completed, templates are thoroughly tested.

3.3 Data capture

Once the survey has been placed into the data capture system for production, the process begins by scanning the questionnaires as they are received. This scanning process produces digital images of each page of the questionnaire. The system uses the developed template to locate the area on the image where field data is located and uses its recognition engine to attempt to interpret the data. During this process, the system is evaluating how confident it is in its ability to correctly determine the data values (those with a low confidence are often referred to as “questionable” fields or characters).

Fields identified by the data capture system fall into two categories; Automated Data Entry (ADE) and Keyed From Image (KFI) fields.

ADE fields are fields that are automatically captured by the system because the system had a high confidence level (above the pre-specified threshold value) when it attempted to interpret the data. In most cases these fields are Optical Mark Recognition (OMR) fields (e.g. checkboxes).

KFI fields undergo a *heads-up capture* operation performed by operators verifying fields on a computer screen. There are three reasons that a field may be KFI; fields that are pre-selected for keying regardless of the system’s confidence level (e.g. 100% verification/key fields), fields where the recognition confidence level falls below a pre-set threshold value established for that field (e.g. data KFI capture), and fields where the system was highly confident of its interpretation but fail some specific edit rule (e.g. range check).

At each of these stages of production processing a potential for error exists. In the case of the automatic data capture of ADE fields, *substitution errors* can occur where the system may interpret a different value from the actual value for that field. These errors are serious since they can be systematic and can significantly impact the overall quality of the data capture process. It is important to be able to track these errors and eliminate them in an effort to optimize outgoing data quality. In the case of the heads-up KFI capture process, *keying errors* can occur where operators may misinterpret the data and may produce higher levels of errors (e.g. manual keying errors).

3.4 Quality control (QC)

The QC process is employed to ensure a given quality level is achieved during the data collection process. This involves using continuous measurements to make decisions on the processes, responding to issues that are observed and making comparisons against quality standards. Survey clients often request assurance that certain data collected is completely error free. These specific data fields are identified as key fields since they are used for identification, aggregation, or critical survey analysis. All key fields in this QC data verification process will have a 100% verification to ensure this error free requirement.

On the other hand, all other data fields not considered as key fields, are sampled, verified and evaluated to ensure that keying and substitution errors are within acceptable limits for the questionnaire. To meet all of these requirements, a two stage QC procedure was implemented.

The initial pass of data capture, operators verify all key fields and all questionable fields (e.g. fields that the system had a low confidence in interpreting the data values). In the second pass, the first stage of QC data verification is performed. A different operator will re-enter all key fields without seeing what was previously entered, resulting in a blind double key verification. If the two consecutive entries are the same the value is deemed correct. If there is a discrepancy between the values entered in the first and second passes, the fields are routed to the second stage of QC where the values are reviewed, confirmed or corrected.

This second stage of QC, also known as the audit phase, consists of sampling and verification of all other fields. The QC procedures are based on the Statistical Acceptance Sampling approach. Under this approach, a separate sample is required to monitor the quality of the ADE and KFI fields independently.

The ADE sampling rate is set based on the survey and does not change. The KFI sampling rate is set based on the survey as well as individual operators and will change throughout the survey cycle. Substitution and keying error rates are recorded for fields within the sample and are used for the following two purposes:

1. To set individual operator sampling rates which can be adjusted weekly based on their performance. These parameterized rates allow for greater flexibility and timeliness in adjusting to keying errors.
2. To evaluate using control charts by comparing the error rate to their respective upper control limits (e.g. rejection levels). These limits are computed by the system based on the expected quality target established for each survey and the sample size of fields selected within the batch. A decision is then made on the acceptance or rejection of the sample relative to the expected quality standard for that survey.

4. Drawbacks of previous system

The previous system used for quality control had a number of significant drawbacks that necessitated the need for a system redesign. From an architecture and implementation standpoint, the previous system was more complicated and required more resources for development, testing, and maintenance. The system required a second set of questionnaire templates developed with custom system coding to ensure that the QC verification would recapture the fields required for verification. The development of these two sets of templates per survey also required that any modifications to templates or scripts had to be made to both sets to ensure consistency of the two processes as they need to be identical in design, in order for the QC to compare accurately.

In addition, the entire QC process was a separate system. Questionnaires needed to be processed through the production data capture steps first, then processed again (following a sampling routine) through the QC system. The comparison of data values was performed only after both production and QC processes were completed and output files created.

The QC sampling strategy that was employed was skip batch sampling with 100% verification (e.g. every Nth batch was selected for 100% verification). Quality decisions of acceptance and rejection were made on the sampled batch. Rejected batches required a manual review of all batches skipped prior to and following the rejected batch to determine whether these should be rejected as well.

Survey clients received their data from the production data capture system while the corrected data was output to a separate file and was only sent to the client when requested. The quality estimates produced from the system only allowed for the incoming error rate estimates to be calculated. In view of the fact that no corrections were made to the errors found, this error rate was estimated to be the same for the outgoing error rate.

Finally, if at any point a batch needed to be investigated, it required many manual steps to track the problem as the data capture and QC processes were not integrated in one complete system.

5. Implementation of quality control

As our software was upgraded to a newer version of OCR for AnyDoc, it included a new feature called the Auditor Module. This module provided an extra layer of quality review, and the ability to selectively define monitoring. It also provided a database layer that allowed the storing of changed data values during the data capture process.

The Auditor module is comprised of three main components: audit criteria, audit phase, and logging. The audit criteria relates to the setting of information for the audit verification step. The audit phase is where data fields flagged for audit are reviewed and corrected if necessary. The logging component stores all field data and process information gathered during the data capture and auditing stage.

5.1 Audit criteria

The audit criteria consists of setting operator sampling rates at the survey level and specifying which data fields to audit at the page level. The acceptance sampling plan for most surveys consist of batch sampling, where for each individual batch a sample of randomly chosen fields are selected for inspection as opposed to skip batch sampling where a sample of batches are selected based on the pre-determined skip batch frequency and the remaining batches are skipped without any inspection. Batch sampling is preferred because all batches are unique and the field-level sampling produces a more representative sample.

Currently the built-in functionality of the Auditor module only allows for skip batch sampling or systematic sampling across documents (i.e. every Nth questionnaire). Customization using Visual Basic Scripting is therefore required to perform batch sampling. To achieve this, three methods were developed by Statistics Canada in order to control the field sampling: verification and control flags; survey specific tables; and QC sampling.

5.1.1 Verification and control flags

Verification and control flags are supplementary fields added to the document template that track various indicators used in the QC process.

Verification flags track three key attributes per field: stage, sample selection, and field value. A verification flag is assigned to each of these combinations, as demonstrated in Table 5.1.1-1.

Table 5.1.1-1
Verification flags

Verification flag	Stage	Sample selection	Stage field value
0	Extract	ADE	Non-blank
1	Pass 1	KFI	Non-blank
2	Extract	ADE	Blank
3	Pass 1	KFI	Blank
4	Pass 2	KFI	Non-blank
5	Pass 2	KFI	Blank
6	Auditor	ADE	Non-blank
7	Auditor	KFI	Non-blank
8	Auditor	ADE	Blank
9	Auditor	KFI	Blank

Control flags assigned to each field determine the action to be taken by the application, as detailed in Table 5.1.1-2.

Table 5.1.1-2
Control flags

Control flag	Action type	Description
0	In sample	Field may be sampled if selected
2	Key field	Field will be 100% verified (key field)
3	Field not eligible	Field not eligible to be sampled (e.g. comments fields)
4	Ignore field	Field will not be verified in any pass (internal fields)

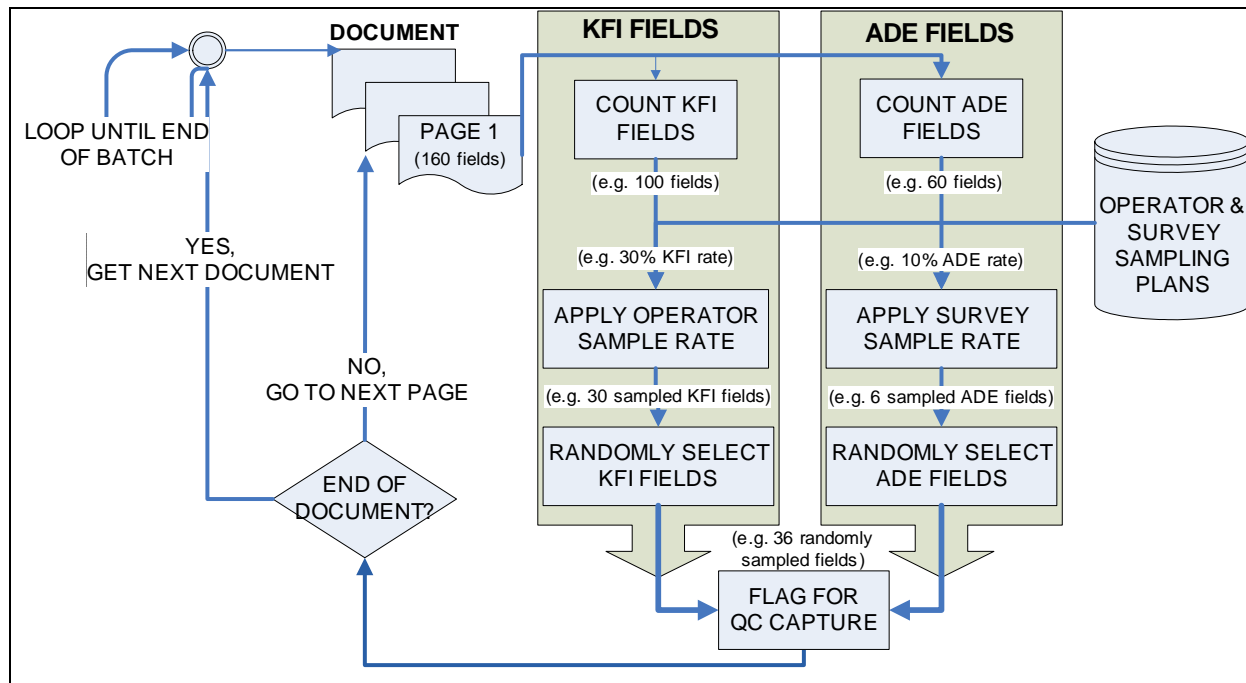
5.1.2 Survey specific tables

Survey specific tables are stored in a database and contain QC data pertaining to individual surveys. QC error targets for ADE and KFI specified in the tables are used as the centre line within three standard deviations, to determine if a batch error rate is accepted or rejected. In addition, the tables contain the operators' sampling rates by survey which are referenced during the sampling process.

5.1.3 QC sampling

After the verification pass, the QC process consists of randomly sampling a number of ADE and KFI fields per page based on their respective ADE and KFI sample rates stored in the survey specific tables. Only fields with a control flag equal to 0 (in sample) are considered for sample selection. The fields flagged for audit will be shown to a keyer during the audit pass. This process of QC Sampling uses the Bernoulli sampling methodology and is demonstrated in Diagram 5.1.3-1 which provides examples for clarity.

Diagram 5.1.3-1
Quality control sampling



5.2 Audit phase

The audit phase pertains to the actual phase where fields flagged for audit are reviewed and corrected, if necessary. Once QC sampling has occurred and the second pass for 100% verification of key fields has been completed, the audit pass begins to display the sampled ADE and KFI fields to an operator for verification. In addition, any key fields where a discrepancy exists between the first and second pass values are also re-verified a third time in this audit phase.

5.3 Logging and QC statistics generation system

Once a batch has completed the audit phase, it is committed and the audit data is stored to a database. Further analysis is performed by the QC Statistics Generation System, an application developed by Statistics Canada, which analyses the audited fields, extracts and tabulates QC statistics to determine whether the batch is accepted or rejected. If rejected, the data for the batch is deleted and the batch is re-captured.

The system has the following functions and capabilities:

- Ability to differentiate (and count) ADE fields from KFI fields and their respective sample size for each batch.
- Ability to identify and count any field as a key field for 100% verification.
- Full record keeping of QC statistics by batch for all field groups (e.g. ADE, KFI, Key Fields).
- Allow the specification of quality error targets for each survey based on the expected mix of field types that will be captured for that application.
- A sample selection module which allows the specification of desired sampling rates by operator and survey for KFI sampling and by survey for the ADE sampling.
- Flexibility to select different modes to sample based on ADE and KFI fields that are either blank or non-blank.
- Comparison of data from the QC audit phase with capture data from Pass1/Pass2 and calculation of substitution and keying error rates by batch.

- Calculation of Upper Control Limits per batch and perform online Control Chart analysis to identify samples that have been rejected and require attention. A listing of rejected batches is produced daily.
- Report outputs containing batch administrative and QC statistical information are created by batch and by field group for the purpose of additional analysis. These outputs are used as inputs into a separate Statistics Canada developed system, Quality Control Data Analysis System (QCDAS), for analysis and to produce estimates and report feedback (to the error source) to the operations area.

6. Benefits obtained with new auditor module

The implementation of a new version of OCR for AnyDoc using the Auditor Module feature and the development of Statistics Canada's own tools and customization has helped to improve the overall QC process. The most notable improvements were: improved sampling strategy; reduced need for development resources; improved QC verification; and improved data quality and quality estimates.

The current system now has the flexibility to allow for a variety of sampling units, such as at a document-level, template-level and/or field-level. In practice, we are sampling at the field level within each document and template for all of our surveys. This permits better QC sampling design which enables us to meet the various survey client needs and ultimately results in better control with minimal additional resources.

The improvements from a questionnaire template development point of view have been substantial. Currently, only one set of templates for both production and QC is needed, therefore greatly reducing the development and maintenance effort and costs, not to mention the potential for errors. Modifications to a template only need to be done in one place and the overall development time has been reduced significantly.

The new system is less complex and easier to implement and maintain. It can be applied to any type of survey, regardless of size or time constraints. In addition, the use of parameterized statistical measurements permits objective process decisions for quality improvement to be made dynamically throughout the production cycle.

The QC verification process is now integrated with the production system as one complete process. By having the QC process in-line, it allows the system to compare the QC values with the production values and determine if differences exist immediately. This in-line comparison allows the QC operator to review values entered during the capture stage, allowing them to self-correct or keep the values originally entered. This self-correct feature eliminates any errors made by the QC operator and ensures that only true errors are recorded. This improvement has greatly reduced the number of discrepancies (e.g. typo and interpretation errors) between production and QC, reduced the number of false batch rejections, and in turn, reduced the amount of resources required to investigate rejected batches.

In addition, there have been benefits to client data quality and the quality estimates produced. The present system replaces the production data with the corrected data from QC. Clients now receive all corrections for data which passes through the QC process. This makes it possible to estimate the incoming and outgoing error rates for any given client. The flexibility and improvements to the QC sampling strategy, the accept/reject decision process and the in-line QC verification have benefited the quality estimates by allowing the QC process to now output true errors as opposed to the discrepancies in the previous system.

7. Conclusion

We have successfully implemented this QC process as our principal method of QC used with the data capture process of paper surveys. It has helped to improve our process timeliness and improve the data received by our clients. In the previous system the process would take approximately two to three days for a batch of questionnaires to be completed and data sent to the client. This improved process has reduced this time to approximately twenty-four hours (1 day). Though the current system has improved substantially, there are still a number of outstanding enhancement requests that have been submitted to the software company AnyDoc. We will continue our efforts to maintain a working relationship with them in order to develop an ideal QC program such as the integration of the

functionality we have developed in the QC Statistics Generation System as possible features in future versions of AnyDoc. We believe our QC process is of sound design and we hope that we will be able to influence AnyDoc in adopting our QC methodology into their software.

References

Mudryk, W. and Xie, H. (2004). Generalized Quality Control Approach for ICR Data Capture in Statistics Canada's Centralized Operations, *2004 European Conference on Quality and Methodology in Official Statistics*. Wiesbaden, Germany.