

## Article

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

### **iCADE the Data Capture System of the U.S. Census Bureau**

by Stephanie Studds

2009



## **iCADE the Data Capture System of the U.S. Census Bureau**

Stephanie Studds<sup>1</sup>

### **Abstract**

In a multi-divisional initiative within the U. S. Census Bureau, a highly sophisticated and innovative system was developed and implemented for the capturing, tracking, and scanning of respondent data that implements Intelligent Character Recognition (ICR), Optical Character Recognition (OCR), Optical Mark Recognition (OMR), and keying technology with heavy emphasis on error detection and control. The system, known as the integrated Computer Assisted Data Entry (iCADE) System, provides digital imaging of respondent questionnaires which are then processed by a combination of imaging algorithms, sent through Optical Mark Recognition (OMR) to collect check box data, and automatically collect and send only write-in areas to data-keying staff for the data capture process. These capabilities have produced great efficiencies in the data capture process and have led to a novel and efficient approach to post-collection activities.

### **1. Introduction**

The iCADE system was created in-house by United States Census Bureau employees with a long history of experience in building large scale data capture systems that are specifically designed to meet demanding census requirements. This system performed data capture on the 2002 and 2007 Economic Censuses. The Economic Census is an extremely complex process consisting of 4,500 distinct page designs, 20,000 data concepts, and over 125,000 answer zone locations that must be recognized and processed by the imaging system. The 2007 Economic Census was processed at 40 percent of the cost of the same census 10 years earlier, and in less calendar time, largely due to efficiencies realized by iCADE.

The iCADE system has significantly reduced the amount of time that employees spend in the National Processing Center (NPC) tracking, handling, and keying respondent questionnaires on a daily basis as well as providing a reliable and efficient solution to managing the workloads by providing the survey sponsors with up-to-the-minute Management Information System (MIS) Reports. After reviewing the results from the Commodity Flow Survey, Company Organization Survey, Annual Survey of Manufactures, Economic Census, Agricultural Census, Private School Survey, American Community Survey (Decennial Census Long Form), School and Staffing Survey, and the Survey of Business Owners, it was determined that our organizations have saved up to 60 percent of their data capture costs in moving from the current key from paper environment to the iCADE System. The cost of iCADE coupled with the storing of the paper image on average for new survey sponsors has been lowered to 19 cents per image.

The efficiency and effectiveness of the system in processing document images have transformed the data capture process from what was once considered a single one-time event. That is, it costs very little to pass through the image files more than once. An initial pass could capture just part of the information that might be required to drive an early, high-priority operation in the census or survey. Then, data that is needed in later operations could be captured in subsequent passes over the images. This is a valuable ability where driving multiple expensive field operations that have varying priorities is important – allowing the up-front operations to receive their information earlier from a partial high-speed data capture pass.

iCADE is also flexible and “programmable” in the sense that the page image “templates” and data descriptive “meta-data” is prepared and fed to the imaging data capture system. From these templates and data characteristic descriptions the system generates the required scanning algorithms and data verification edits. A measure of success of this design is that the Economic Census, the Agricultural Census, the American Community Survey, the

---

<sup>1</sup> Stephanie Studds, United States Census Bureau, 4600 Silver Hill Road, Suitland, MD, USA. 20746

Commodity Flow Survey, the Survey of Business Owners, the Business R&D and Innovation Survey, and a number of smaller censuses and surveys – all with very divergent requirements – were simultaneously processed with a single system – for the first time in U.S. Census Bureau history. The generality and ease of use of this system is such that we will be adding 10 to 15 new data capture applications annually.

The system uses a combination of document based Optical Mark Recognition (OMR) checkbox analysis, high-speed Key from Image (KFI), and automatic recognition of hand-written characters (ICR). The imaging system document processing and visual presentation is lightening fast, even on the most modest of computer platforms. This is a single document imaging software application that performs the document registration, ICR, OMR, keying, verifying, classification, and adjudication. The software presents survey page images with captured ASCII data spliced into the visual presentation, along with error warnings, keyer PostIt-like notes, and status flags describing respondents' marginal notations. The same software and presentation, subsequently, is used for the survey sponsor's review and for historical review (historical review is valuable for questionnaire design and response rate studies). All generations of the OMRed, ICRed, and keyed data are preserved and presented for review. Nothing is ever deleted or overwritten – so a complete and open review or analysis of every step of the data capture interpretation process can be performed (and is encouraged) at any time. (The document images and captured data are available to analysts in a single visually attractive presentation throughout the data capture process, including years after the completion of work.)

The system also contains an exacting "Control System", which controls every batch through the system. The Control System's responsibilities include: process control, tracking, quality control (QC) measurement, interpretation error containment and control, and extensive work progress, performance, and data quality reporting functions.

## **2. Sponsor planning for migration to iCADE**

Surveys migrating from the key from paper environment to the iCADE System require approximately 6-10 months of planning, designing, and testing prior to production implementation. The first phase of planning includes scheduling the project from start to finish in Microsoft Project Software, the set-up of frequent planning/status meetings throughout the life-cycle of the project, identifying the "final decision" maker for the survey, discussion of monetary resources, and the completion of the Project Registration Package. Phase two of planning includes information from the survey sponsor on the number of questionnaires to be processed and how and where the data will be reformatted and delivered to the sponsor to determine the necessary IT and programming requirements. The third phase of planning includes the reviewing of the survey questionnaire and the number of questionnaires expected to be returned from respondents in conjunction with current staff in NPC for the project to determine future staffing needs for Check-in, Open and Sort, Batching, Scanning, Manual Registration, Exception Review, Keying, and Quality Assurance personnel.

The survey sponsor is also responsible for meeting with the iCADE Team to determine project requirements such as the sample verification sample rates, sample verification allowable error rate, business rules for the Batching Process, business rules for the Exception Review Process, business rules for reformatting and data output, business rules for production batch resets, and the business rules for any post data capture/pre-data output Analytical Review Process.

## **3. iCADE system**

### **3.1 Forms design**

The first step in the migration process to the iCADE System is forms design. The forms design process provides key support for iCADE which is the metadata. The survey questionnaires are designed utilizing the KFI Specifications for answer zone creation for the iCADE process, which can be done in either the AMGRAF ONEFORM software in the National Processing Center or in the Economic Metadata Repository (EMR)/Generalized Instrument Design System (GIDS) at headquarters. The AMGRAF software requires a trained

individual to design the forms and requires little metadata input on the front end of the forms design process allowing harvesting of the x and y coordinates from the XML component for the data capture process. The AMGRAF software requires an analyst to compile all of the metadata necessary for the iCADE process such as Data Element Names, Data Concept, Quality Assurance Requirements, Character Set, Capture When etc. and place it manually in an excel spreadsheet for the creation of the master template. On the other hand, the EMR/GIDS software tool can be used by non-programmers to create complex survey questionnaires, which can be deployed into paper and electronic instruments for survey response collection. The EMR/GIDS software requires the metadata be defined during the forms design phase and automatically produces the excel spreadsheet format done manually in the AMGRAF process.

Metadata determined during the forms design process includes legal character sets and value range edits for each applicable field. The metadata also indicates whether the write-in field should be captured, presence detected, or skipped. "Capture When" is set to "C" when capturing/keying the field. "Presence Detected" is set to "P" when a field has presence but will not be keyed; a "P" will appear for the field in the output file. "Capture When" is set to "S" when the answer zone is to be skipped, even if the response is given; the field will not be presented to the keyer for data capture and therefore will not be in the output file. Finally, the survey sponsor has the ability to provide legal code lists and/or drop down auto-complete listings that will be utilized during the Data Capture Phase to validate items keyed such as Cities, States, Zip Codes, Commodity Codes, North American Industrial Code System (NAICS) codes, etc.

### **3.2 Batching**

The Batching process allows millions of questionnaire forms into the National Processing Center on an annual basis to be placed into a logical sized work unit called a batch. A batch can contain any number of forms, which is sponsor specific and decided early on during the planning phase. The system is sophisticated enough to recognize the integrity of the work units and track them throughout the life cycle of the process. The intelligent program utilized by the batching process automatically reads the 22-digit barcode on the front page of each survey form which is validated against the mail universe and creates the form batch per the survey sponsor specifications.

### **3.3 Form preparation**

During the Form Preparation Phase, 11x17 booklets bound together with spine staples are sent to the guillotine where the spine and staples are removed prior to scanning. The 8 ½ x 11 forms with staples in the upper left corner of the survey forms are removed by biscuit cutters.

### **3.4 Scanning**

In the Scanning Process, batches are entered into the iCADE System and are scanned using the KODAK 9500D to produce Group4 compressed .tiff images scanned at 200dpi resolution. The programmers have integrated the iCADE System with the Kodak scanning software into the control system by using Dynamic Linked Libraries (DLL) to run JAVA through the JAVA Native Interface (JNI) to communicate the scanning activity to the control system. The scanning software then transfers the images to the SAN and the workflow information necessary to proceed to the next phase of processing.

### **3.5 Auto registration**

Auto-Registration is performed on servers under automatic control. In this phase, the system examines processed images and automatically matches them with the unique page templates from the Master Template and the page barcodes imprinted on the forms. This phase is also where the OMR checkbox answers are read, evaluated, and captured and the presence of respondent hand-written entries are detected. It is also this stage where ICR is performed on all specified numeric write-in fields. Presence detection is performed on all "PresenceIsAnswer" fields and the results are stored in the script file. The document skew is determined and correction factors are recorded in the script file for future image display correction. The questionnaire ID barcode is read and recorded in the script

file. Any questionable “field barcodes” are interpreted and stored in the script file. An initial “script file” is produced and will be used to control the next task of Manual Registration.

### **3.6 Manual registration**

The Manual Registration phase allows for forms that fail auto registration to be manually registered by four corner point digitizing of the form. This process also presents questionnaires and page type barcodes that could not be interpreted, and pages which could not be automatically registered to be presented to a manual registration clerk to identify the page corners or manually key page barcodes. This process allows damaged pages to be recovered and repaired from the images – eliminating the labor-intensive need to retrieve, repair, and rescan the damaged paper forms. This process, also allows the “Got Presence” flags to be reset and the OMR the opportunity to reevaluate the fields for presence detected (non-blank) and no presence detected (blank).

The Manual Registration process for the OMR presence detected (non-blank) occurs in two phases. In the first phase of Checkbox Ambiguity Repair, checkboxes with more than one response selected are reviewed by a Manual Registration clerk. This applies even if the survey sponsor allows for “more than one” answer within a series of checkboxes. Since checkboxes are never presented to a keyer for data capture, this is the only phase where answers will be evaluated for their validity. It is common for respondents to extend the “x” or “checkmark” tails from the marked answer zone into a non-marked answer zone. The software evaluates the extraneous mark to see if it meets the criteria for the signature of a valid mark inside a checkbox. If it does meet the criteria and there are more than one in the answer series for a question, the response is sent to a manual registration clerk for review.

The other process in Manual Registration is “Noise” Processing. This is used currently by the Economic Directorate as one additional review of “blank” and “non-blank” checkboxes on the survey questionnaires. New postal equipment being used by the United States Postal Service has the belts on the equipment set to a very high tolerance for non-flat envelopes to pass through the equipment to minimize jams and maximize flow efficiency. This process can cause ink to come off of the survey forms and fit the criteria of a response signature in the checkbox area of the survey form. As a result, the “Noise” software was developed as a two-fold review process. The first step looks at non-blank fields. The manual registration clerk is presented on one monitor approximately 360 checkboxes per page from forms in the batch, while at the same time, on a second monitor (as the clerk pans across the checkboxes) the full image is being displayed for review. The clerk can at this point de-select or select a response if it has been incorrectly or not marked by the system. The second step in this process is the review of blank fields. This part of the process presents checkboxes that are believed to be blank on one monitor and as the clerk pans the checkboxes, the clerk is reviewing the full image of the form on the other monitor. This allows the clerk to select the fields that should be non-blank and were believed to be blank.

### **3.7 Exception review**

As the forms exit the Manual Registration phase of the system and enter the Exception Review phase, a series of Batch Completeness Procedures and validations are done. During the earlier Batching phase, the wand of the barcode indicates to the system how many pages a particular form should have. If a form is identified to have missing or loose pages that do not belong to the form, they fit the criteria for the Exception Review process and are written out to a listing for review by an Exception Review Clerk. This is also the location where “Batched but Not Scanned (BNS)” and “Scanned but Not Batched (SNB)” cases are identified. The BNS cases are wanded into the batch during the Batching Phase but are not seen at scanning. The SNB cases are not wanded into the batch during the Batching Phase but appear during Scanning. These scenarios are also written out to the Exception Review listing to be reviewed by an Exception Review Clerk. The Exception Review Clerk receives the paper batch in the batching container with the exception review listing at their workstation. They review each form on the listing in conjunction with the survey specific requirements and must resolve each form in the batch by assigning resolution codes in the system.

If the resolution code set by the Exception Review Clerk is “Unresolved”, then the ID is removed from the script file and the form is sent back to the Batching Phase to be re-batched. If the cases are resolved with “As Is” codes, the ID and script records will remain in the batch for data capture. At this time the fields are also flagged for “Key from Image”, the next step in the process. In the metadata process for the creation of the Master Template, the survey

sponsor indicated which fields should be “detected for presence” and data captured. Those fields are having their “Capture When” set and their “Capture When” values output. The “Capture When” is set in conjunction with the Data Concept from the Master Template. The values are also assigned for “Multi-Pass” keying at this time which allows the keyer to key all numeric fields in one pass thus maximizing the use of the number pad on the keyboard and then alphabetic only and then alpha/numerics together.

This is the most critical phase of the iCADE process. It is here that a human can determine if the forms printed were not printed correctly, if a respondent has done creative scanning in their response of the form, or if the forms have been placed into the scanner incorrectly creating what the system refers to as a “train wreck”. A human can intervene to pull these forms prior to keying and output to the survey sponsors.

### **3.8 Key from Image**

During the KFI Process, respondents’ hand-written data is presented to a data keyer using the metadata from the survey’s master template by page type. Numerous edits are performed on each field that is keyed. Validation edits are used during this phase to enhance the keyers’ performance and minimize data entry errors. A keyer may also “force” into a field where “Presence” was not detected as having write-in information because the respondent wrote the answer outside of the answer box. The keyer has the ability to suspend work on a batch and return to it later. The system automatically indicates to them where they left-off when they return to the batch. The keyers see the entire page image, and they can also view the whole multi-page document and, therefore, can consider the context when unusual response interpretations are required.

The fields that are set to no “Capture When” will not be presented to a keyer. During the initial keying phase, a keyer can force into fields without “CaptureWhen” set to key answers which were entered in the wrong location by the respondent. After initial keying is complete, fields without a “CaptureWhen” will not be able to have data entered for them in subsequent keying tasks.

Key-from-Image is preferred over automatic ICR recognition by some response field types for six reasons: First, many respondents of census surveys are of advanced age and survey sponsors would prefer to have human keyers interpret the respondents’ free-form writing than to have the respondents conform to machine induced requirements (thereby reducing respondent burden, increasing response rate and level of quality to the data captured); Second, the keying presentation allows the keyer to flag or code marginal information (for example, respondents change the questions asked or provide multiple answers for a single answer zone); Third, because ICR fields require a dictionary or totaled columns as an accuracy check – where these are absent keying is preferred; Fourth, because many casually implemented ICR systems are grossly inaccurate; Fifth, because in the experience of the U.S. Census Bureau, the cost of keying is less than 20 percent of the cost of large scale contractor-supplied ICR systems; and Sixth, in order to be accurate, ICR systems have to be backed up by extensive keying repair operations that frequently are less efficient and less accurate than straight-out high-speed keying in non-repair mode.

### **3.9 Quality assurance of OCR fields**

The Quality Assurance Phase is split into Verification, Classification, and Adjudication for respondent write-in fields. Once the batch has completed the Key from Image phase, the batch enters the first segment of Quality Assurance. It is based on the Quality Assurance specifications set by the survey sponsor for the project, that percentages of blanks and non-blank fields are set in the algorithm for what they would like selected for Verification Keying.

The system has been meticulously designed and implemented to provide accurate interpretation of respondent’s written answers to census or survey questions; and to provide the best available quality assurance mechanisms to rigorously measure and control the error inherent in the interpretation process; and to accurately control and track workflow (down to the individual page level) – allowing no work to be duplicated or lost; and to provide detailed reporting capabilities on work flow tracking and status, machine, and personnel performance and accuracy, and work backlog and degree of completeness measurement.

After ICR or first-pass keying-from-image, a sample of the fields (usually a 3 to 5 percent sample) is selected by the Control System and presented to an independent “verifier” keyer (who cannot see what was ICRed or keyed by the first-pass keyer). The Control System then uses the sample to generate an error rate for the batch, by keyer. Fields that have errors are sent to an Error-Classification process and then to an Adjudicator process. Batches that exceed a defined error threshold are reprocessed. Accuracy and productivity measurements are accumulated for each keyer, and for the ICR process. Error measurement and control is a heavily emphasized function in this data capture system. The census or survey sponsor can regulate the quality assurance sample size (up to and including 100 percent) and therefore control the interpretation error to any level desired. Detailed error reports provide valuable feed-back to keyers, procedure writers, trainers, survey designers, forms designers, and to survey sponsors.

### **3.9.1 Verification and classification**

In Sample Verification, a different keyer is given the selected sample of fields to be verified and independently keyed. Any field not selected for verification will not allow new values to be keyed for them. The selected sample field answers keyed by the first and second keyers are then compared. If there are any differences in the keying, the fields are then identified for Classification, the next phase in the process. During Classification, a third keyer is selected to provide a classification code for each identified difference from the Verification phase. The error codes must be assigned for each discrepancy in accordance with the procedures provided to the classifier and the Quality Assurance Auditor. By assigning the correct error codes, the system ensures that only good batches are accepted, the rejected batches are rectified, and the proper data is retained for output to the survey sponsor. The system is programmed to distinguish between chargeable and non-chargeable errors based on the codes assigned by the Classifier. It is the primary function of the Classifier to assign error codes; however, if the Classifier determines that both the keyer and the verifier are incorrect, then the Classifier will have to provide an answer string for the data item.

### **3.9.2 Adjudication**

In the Adjudication phase, the fields Classified from the earlier process are sampled and Adjudicated by another independent group. The Adjudicator can agree with the keyer, agree with the Verifier, agree with the Classifier, or provide their own interpretation of the respondent’s answers. The Adjudicator will provide a classification code for all of the differences found during this process. As a result of the Adjudicator’s review and all previously classified differences, a decision is made on whether or not to pass a batch. If it is determined that the batch requires remainder processing, all fields not selected in the original verification sample will be selected for “Remainder” processing. The sample rate may differ between the blank and non-blank fields but will be consistent across OMR and KFI fields. If a batch receives a “Reject” decision and the Quality Assurance Plan calls for verification of blanks and non-blanks, then the remainder verification will include ALL remaining fields within the process (OMR or KFI) regardless of presence detected (non-blank) or no presence detected (blank) and independent of why the batch failed. If only non-blank fields are selected for sample verification, then only non-blank fields will be eligible for remainder verification. The same will be the case for blank fields.

### **3.9.3 Remainder verification**

In Remainder Verification a keyer different from the original keyer will be directed to the fields selected for “Remainder” processing. This keyer will not be able to see what the original keyer keyed. Only the fields selected for Remainder Verification will allow answers to be keyed for them. The answers keyed for the selected fields are keyed and then compared against the original keyer’s data. The fields that contain differences are selected for the classification process. The remainder batches are tracked by the keyer on a daily basis and, based on survey sponsor criteria, a keyer’s qualification can be impacted and the keyer can be removed from the data capture process. The system will compute differences in all of the fields, the types of differences, and the error rates. Based on the Adjudicators pass/fail recommendation, the batch will either be forwarded for final processing or will be repaired.

### **3.10 Analyst review**

The Analyst Review process is optional and can be set-up with special code written to enforce the business rules specified by the survey sponsor, or a manual flag can be set by a keyer in the Key from Image phase that sends the questionnaire to post processing review. Depending on survey specific requirements, the form may be selected for review by the system code or by a keyer and then, when it is reviewed by an analyst, it may be removed for a batch. If the questionnaire is to be removed, the remaining questionnaires will be sent to the Output phase and not be held up awaiting final results of the review process. However, if the survey sponsor decides that the integrity of the batch must be maintained and the entire batch must be released at one time, the time required for the review has been built into the Microsoft Project Schedule for delivery.

The survey sponsor may designate a final review team consisting of analysts for the sponsoring division or agency. The authorized team members are given read-only access to the post-processing output files via a system of search screens. Reviewers can search for data items that were changed in Analyst Review, for specific forms in a batch or for specific batches. If the analysts determine that changes should be made to the data before it is delivered, a request is made through survey specified guidelines set-up during the planning phase.

### **3.11 Output**

During the Output phase, a daily pooler file of completed batches is reformatted and edited as per survey sponsor requirements and loaded to the survey database.

### **3.12 Feith document database**

During this phase as we pooler the output files, the tiff images of the survey questionnaires are sent to the Feith Document Database (FDD). As the survey analysts are performing micro and macro-level review of the survey data, they can review the survey questionnaires in FDD as well as the keyed data in the database. Based on survey specific requirements, the FDD index and tools can be set-up for interactive updates of correspondence, undeliverable as addressed (UAA), and respondent contact information by the survey analyst.

### **3.13 Management information system reports**

The Management Information System (MIS) Reports were designed in a general manner to facilitate providing real-time information for any survey utilizing the iCADE System. The system contains production as well as keyer-level reports. This assists the National Processing Center as well as the survey sponsors in making effective decisions with the use of resources. The system has significantly reduced the amount of time that employees spend tracking, handling, and keying respondent questionnaires on a daily basis as well as providing a reliable and efficient solution to managing the workloads.

## **4. Summary**

After reviewing the results for the Economic Directorate Current Surveys, the American Community Survey, Economic Census, and the Agricultural Census, it was determined that our organization has saved up to 60% of their data capture costs coupled with a less than 1 percent error rate in moving from the key from paper to Key from Image technology. Since the iCADE system was placed into production, it has far exceeded the initial goals including (1) utilizing the XML outputs from the AMGRAF forms design system to define fields and using graphical inputs from pdfs, jpgs or tiffs to assign metadata necessary for the Manual Templating Process for each survey (2) continuing to enhance the inter-field editing capability and dynamically linked libraries, which allows for more real time data editing during the data capture process (3) building one comprehensive system that facilitates the Batching, Scanning, Registration, and Exception Review process, and (4) containing an Analytical Evaluation Module that allows analysts to review data in real-time prior to it going through the output process.

The U.S. Census Bureau envisions this data capture system as the leading edge technology for future data capture of all of its Economic and Demographic paper-based questionnaire processing.