Topic: Field Reengineering

KEY-from-IMAGE:
A tool to enable the devolution of data cleaning phase in document imaging data capture technique down to the lowest level field office (Provincial Office) to minimize "mindless" imputation in automatic data editing

Prepared by Gene V. Lorica, Philippine Statistics Authority, National Statistics Office, Philippines

I. Introduction

1. Ten or more years ago, the use of document imaging technology for data capture in surveys and censuses was the craze in statistical offices where several countries actually tried it including the Philippine National Statistics Office (PNSO). Data capture using Document Imaging technology is a technique where the images of filled-out survey/census forms are captured into electronic format using high or mid-volume scanners producing a scan image file for each form. Data in the scan image files are automatically extracted using a Document Imaging Software (DIS). Specifically, typewritten data are extracted using Optical Character Recognition (OCR), marked fields data are extracted using Optical Mark Recognition (OMR), and write-in data are extracted using Intelligent Character Recognition (ICR) functionalities of the DIS.

2. Conceptually, the use of the technology for surveys/censuses is considered as much more superior than conventional data entry system (key-from-form) for the following reasons:

(a) Data capture is much faster;
(b) Data capture is more accurate (*as claimed by Document Imaging Software vendors*);
(c) The system requires fewer resources although not necessarily less expensive since scanner and imaging software are expensive;
(d) Electronic copy of forms can be kept and made use of for a long period of time.

3. Unfortunately, based on the experiences not only by the Philippines but also by other countries that used it, the document imaging data capture technique was found to be all hype. The main problem of the data capture technique is that the accuracy of the automatic extraction of write-in entries from the scan image file of survey/census forms is far from what is considered as reliable. Because of this drawback, most of the alleged advantages of the technique were actually just a hard sell by the software vendors.

5. As a solution to the incorrect capturing of write-in entries during the 2000 CPH, verification of data was done by encoding those data items using a crude data entry system that runs on a separate window side-by-side with another window showing the scan image of the form on the same display screen – a gargantuan task for over 20 million forms.

4. After the data capture verification, an extensive automatic data editing and imputation was implemented. But during micro level verification of the changes done by the automatic editing, it was found that there was significant number of correct data that were replaced with consistent but spurious entries. Because of this experience, the office believes that there is no substitute in scrutinizing the data in the questionnaire in resolving intra-record and inter-records data inconsistencies found and automatic editing should be minimized.

5. Unfortunately, scrutiny of the data by referring back to the forms in resolving data inconsistencies is a very tedious task and it is deem too difficult and too costly to implement. The solution that the office thought of was to make it easier to link error messages to their respective affected scan image so that the need to refer back

to the forms can be eliminated, that is, a data editor can open/view the scan image and reflect the corrections both in the data file and the image file.

5. It is a known fact that not all is bad for the document imaging data capture technique. The accuracy of automatic extraction of entries in marked fields using optical mark recognition (OMR) is very high with less than 1% error in interpretation. Another invaluable feature of the technique is that the scan image files of survey/census forms can be conveniently kept and made use of for a long period of time. In the case of the Philippines, the scan image files of the census forms were used in the design of the Master Sample and in generating the list of names of household heads and addresses of selected sample households. These are the reasons why PNSO, despite the setbacks experience in 2000 CPH still used this method in succeeding censuses and will continue to do so in the coming 2015 mid-decade census of population.

6. To exploit these advantageous characteristics of document imaging and to be able to transfer the data cleaning tasks to the provincial offices, a hybrid survey/census data capture technique was contemplated by the PNSO. The hybrid technique is a combination of document imaging and data entry application with some special traits. Characteristically, our hybrid system covers the following general steps:

(a) *Scan the forms;*
(b) *Extract OMR data from the scan image of the forms using Document Imaging software;*
(c) *Send scan image files and interpret batch files to their respective provincial office (PO);*
(d) *POs do the data entry, key-verification, and data cleaning using a program that enables encoding data direct from the scan image file- eliminating the tedious task of referring back to the forms during data editing*

7. Thus, the birth of **Key-from-Image (KFI)** program.

II. Evolution of the KFI (Summary of PNSO Experiences in Document Imaging)

A. 2000 Census of Population and Housing

8. General description of the system used:

(a) The system development was out-sourced from Fujitsu Philippines – a private company;
(b) The outsourced system is a VB6 program that combined Mid-Volume Capture Software (MVCS), Document Imaging Software, and some Control, Management, and Report mechanisms;
(c) The system facilitated the use of mid-volume Kodak Scanners with Mid-Volume Capture Software (MVCS) for converting questionnaires into electronic format (scan image files);
(d) Document Imaging Software was used to extract write-in entries using Intelligent Character Recognition (ICR) and Optical Mark Recognition fields (OMR) from scan image files;
(e) Questionnaire has drop out colour, i.e., texts, frames, boxes, and lines are printed in colour that the scanner would not detect, hence, the scan image file contains hand-written texts and marked fields shading only;
(f) Final output data files are in Integrated Microcomputer Processing System (IMPS) file format. IMPS is a DOS-based census/survey processing system developed by the US Bureau of Census;
(g) Data processing was done in 6 data processing stations (semi-centralized).

9. Problems encountered:

(a) Because the system was designed for data capture of documents in general but not specifically for survey/census processing where forms must be processed by folio/batch, it resulted to difficulty in consolidating the data by enumeration area;
(b) Data entry verification had to be done using Centry of IMPS for all extracted write-in data because it was discovered that the supposedly intelligent character recognition (ICR) is not that intelligent as

claimed by the software vendor. The extracted write-in data when compared to verified data shows that the former produced severe 'garbage' results;

(c) Data cleaning could not be delegated to the Provincial Offices which restricted the office to resort to an extensive automatic data cleaning.

B. 2007 Mid-Decade Census of Population

10. General description of the system used:

(a) Used mid-volume Kodak scanners (the same units used in 2000 CPH and 2002 CAF) with MVCS;
(b) Used the same Document Imaging Software but was tailored for census data processing thru ActiveX DLL (Dynamic Link Library) to extract mark fields entries only, and to enable the processing of scan image files by folio/batch – forms were folioed by Enumeration Area (EA);
(c) Instead of using OMR functionality of the software, 'X' marks in mark fields were extracted using ICR whose behaviour was refined by programming event handler through ActiveX DLL.
(d) Developed VB6 Key-from-Image program designed specifically for this census to enable data encoding and key-verification of write-in entries direct from the scan image files. The program also allows correcting extracted and/or encoded data direct from the scan image file without referring back to the hard copy documents;
(e) These programs were integrated in the Survey/Census Integrated Processing System (SCIPS) – a VB6 program designed to facilitate control and management of the data processing of each batch/folio;
(f) Data processing was done in Regional Offices (ROs);
(g) Final output batch files are in Census and Survey Processing System (CS Pro) file format;

11. Limitations:

(a) Since the processing centre was far from the source of the data, quite significant amount of unknown and inconsistent data had to be imputed;
(b) The program had no capability for the user to reflect or 'write' corrections in the scan image file;
(c) Reusing the scan image files required reprogramming the Key-from-Image program.

D. 2010 Census of Population and Housing

14. General description of the system used:

(a) Used mid-volume Kodak scanners - the same units used in 2000 CPH and 2002 CAF) plus additional Panasonic scanners, with new Kodak Capture Software (KCS) that replaced all MVCS;
(b) Used the same Document Imaging Software but was tailored for census data processing through ActiveX DLL to extract mark fields entries only and to enable processing the scan image files by folio/batch. This is basically the same technique as in 2007 Mid-Decade Census of Population;
(c) Key-from-Image program was enhanced to enable 'writing' or reflecting corrections in the scan image files;
(d) Redesigned Key-from-Image program to make it more user-friendly and multi-purpose, so that it can be used in any survey/census and in different types of data capture situations through the use of a configurable template file;
(e) Similar to the 2007 Mid-Decade Census of Population, these programs were integrated in the Survey/Census Integrated Processing System (SCIPS);
(f) Developed/enhanced the KFI Template Editor – a user-friendly program designed for the preparation of the template file for each type of form. This made Key-from-Image program a full-fledged software system;

(g) Scanned the forms and extracted OMR fields in their designated scan station then sent the files to their respective Provincial Office.
(h) Let the Provincial Offices (POs) do the data encoding of write-in entries, key-verification, data cleaning, and evaluation of resulting population counts;
(i) Final output batch files are in CS Pro file format.

D. 2015 Mid-Decade Census of Population

15. General description of the system that will be used:

(a) Will use mid-volume Kodak scanners - the same units used in 2010 CPH and 2012 CAF) plus additional scanners with Kodak Capture Software (KCS);
(b) Will use the same technique as in 2010 Census of Population and Housing;
(c) Coding will be done using the built-in e-CodeBook in Key-from-Image

III. CONCLUDING REMARKS

17. The following advantages/benefits were realized out of implementing hybrid data capture system utilizing the strengths of the Key-from-Image program:

(a) Saved cost on printing of forms since the only requirements are the forms must be printed exactly the same as the template and the paper weight must be at least 80 grams per square meter (gsm).
(b) Data entry of handwritten entries direct from scan image of the forms is much faster than conventional data entry;
(c) Lessened data items for encoding, since OMR data items can be automatically extracted from the scan image files of the forms, with very high level of accuracy;
(d) The technique made it possible to do verification of keyed-in and extracted data items;
(e) Allows corrections to be reflected in both the data file and the scan image file of the forms;
(f) Enables the devolution of data encoding and data cleaning down to the lowest level field office (Provincial Office) which resulted to minimum “mindless” imputation/changes done in the automatic editing phase;
(g) Accuracy of keyed in data items ranges from 99% to 100%, based on sample key-verification in the 2010 Census of Population and Housing;
(h) Enabled PNSO to reuse the forms in preparing a Master Sample, which is a list of sample households, by keying in the names and addresses of respondents directly from the scan image files of the sample household forms;
(i) Because the number of data items imputed or changed during the automatic editing is very low, the office is confident that the resulting population counts and characteristics represent the real-life situation in the country.

18. With insights gained over the timeline that provided for the developmental evolution of the Key-from-Image program, the Philippine National Statistics Office found a reliable and workable solution for the inherent problems in using document imaging data capture technique for censuses and surveys. The Key-from-Image program continues to evolve as it steadily adopts better strategies over data capture using document-imaging technology.

Appendix - KEY-from-IMAGE Software

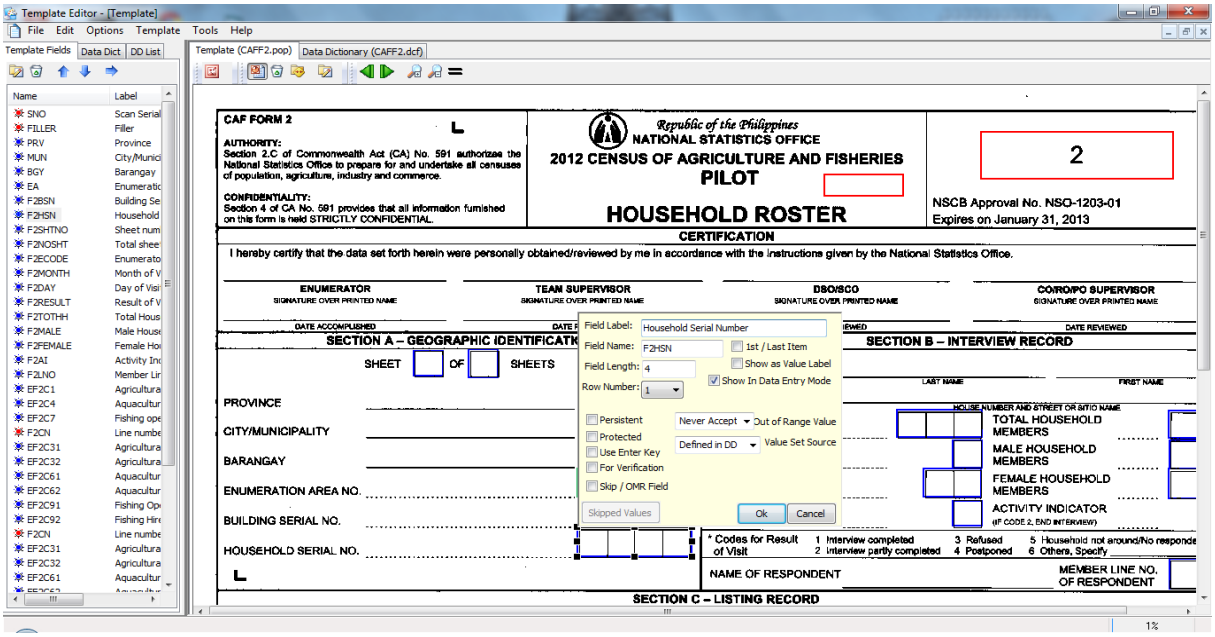
A. Main Features of the Key-from-Image Software

1. The software can be used for different types of data capture techniques. These are:

   (a) Data encoding direct from scan image file with or without previously interpreted/extracted OMR data fields;
   (b) Data encoding without scan image file (or conventional data entry application) where the data screen is exactly the same as the forms;
   (c) Template based data encoding that creates filled-out image file for each encoded form;
   (d) Data encoding with sample or 100% key-verification.

2. Using the software for other survey/census needs no reprogramming the KFI. A template file can be prepared instead using the user-friendly KFI template editor. The characteristics of the KFI Template Editor are:

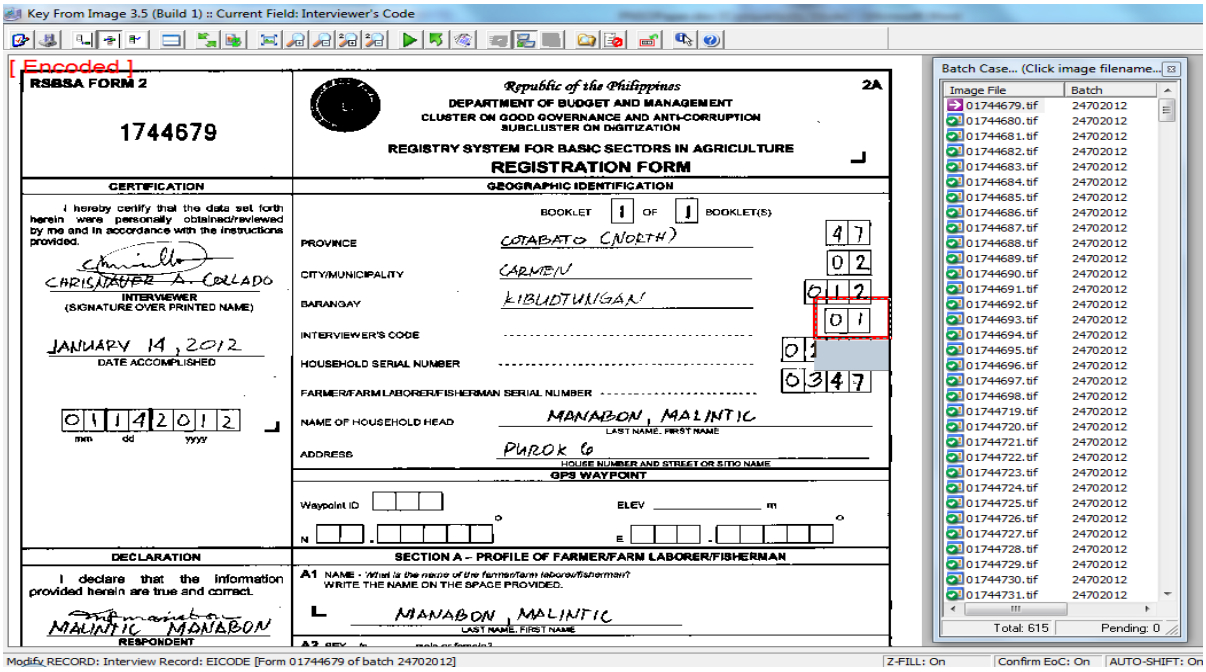
   (a) User can prepare the Data Dictionary either by using either CS Pro or the Data Dictionary module of the KFI template editor;
   (b) Draw rectangles then drag and drop user interface in linking the Data Dictionary items with their respective data items in the template scan image file (empty/not filled up/blank) and specifies the properties and behaviour of the data items during data encoding;
   (c) Skipping patterns can also be specified using the Template Editor.



3. Data encoders who are familiar with CS Pro can easily adapt to the system because some of the commands and shortcut keys are the same. The main features of the KFI run-time program are:

   (a) Intuitive main screen menu options and sections (windows);
   (b) Users can switch to/from “Add”, “Verify”, or “Modify” Data entry modes;
   (c) Program facilitates zooming in and out the main screen or the data window where the image is viewed

(d) The data entry box can be set to different positions and/or behaviour, and these are, by way of enumeration, floating right below, or inside each data cell, or stationary position either at the upper right corner or at the bottom of the window;
(e) The data window (data content of the current position of the data cell pointer), if set to visible, can also be positioned either at the upper right corner or at the bottom of the screen;
(f) Facilitates writing corrections (annotate) in the scan image file.



B. Preparing the KFI Template File

4. The following steps describe the simplicity in preparing a KFI template file.

(a) Create data dictionary using CS Pro (recommended) or the KFI Template Editor;
(b) Scan a blank form (not filled-out) which will be used as a template image file. The size in pixels must be around 1650 x 2400.
(c) Run the template editor, then load the data dictionary and blank image file.
(d) Link each data dictionary data item to their respective data cell in the form by drawing a rectangle then drag-and-drop the corresponding data dictionary item;
(e) Set the data item properties and behaviour;
(f) Set skipping pattern;
(g) Set template application's properties;
(h) Save the template.

C. Data Encoding Using the KFI Program

5. The following steps describe the data encoding using KFI:

(a) Run Key-from-Image program;
(b) If no command line parameters were supplied, the program prompts the user to either use the previous setup or to supply new run parameters;
(c) When the program runs, the user will be prompted to type or encode all data items as defined in the template file skipping OMR fields which have already been interpreted using the Document Imaging Software;

(d) The program checks the validity of the entry based on the defined value sets in the data dictionary or even in a reference file;
(e) It also does simple data consistency validation and checking as specified in the template file;
(f) Follows the skipping pattern specified in the template file;
(g) Users can write/annotate corrections in the scan image file;
(h) Users can locate a particular code for categorical variables with long list of possible values such as Occupation, Industry, etc., using the built-in electronic CodeBook support.

D. Calling the Program from another Program (Integration)

6. The program can be called from another application using the Shell command. It therefore supports parameter arguments passing over the command line:

/UN[*User Name*]
/**DS**[*Data Source Path*]
/**DD**[*Data Destination Path*]
/**EP**[*Encoded Path*]
/**JP**[*Job Path*]
/**TF**[*Template File*]
/**FP**[*For Rescan Path*]
/QS[*Questionnaire ID*]
/DM[*Data Entry Mode*]
/SB[*Settings Button*]
/VR[*Key Verify Rate*]

Where keywords used are:

/**UN** - User Name
/**JP** - complete folder path containing scan image files to be encoded
/**EP** - complete folder path where encoded scan image files will be moved to
/**FP** - complete folder path where unreadable images would be moved to
/**DS** - complete folder path where batch file (output of interpretation) would be read
/**DD** - complete folder path where encoded batch file would be written to
/**TF** - complete template file path
/**DF** - not used
/**QS** – ID of Questionnaire to be opened
/**DM** - set value to 'Add', 'Modify', or 'Verify'
/**SB**
/**VR** - key verification rate [default is 100]

7. Below is a sample command line with parameters passed:

KFI /UNDE1101 /TFZ:\CPH2010F3\CPH2010F3.pop /DSZ:\CPH2010F3\INTERPRETED /DDZ:\CPH2010F3\ENCODED /EPZ:\CPH2010F3\IMAGES\372030030000\Encoded /FPZ:\CPH2010F3\IMAGES\372030030000\Rescan /JPZ:\CPH2010F3\IMAGES\372030030000 /DMVerify /SBFalse /VR50

F. Hardware/OS Requirements

8. The KFI program runs on hardware and software systems that support Microsoft Windows XP as a baseline operating system. It can safely run on systems running on Windows 7 and Windows 8.