UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

UNITED NATIONS
ECONOMIC AND SOCIAL COMMISSION
FOR ASIA AND THE PACIFIC (ESCAP)

Meeting on the Management of Statistical Information Systems (MSIS 2013)
(Paris, France, and Bangkok, Thailand, 23-25 April 2013)

Topic (iii): Innovation

KEY-from-IMAGE:

A workaround for a not so 'intelligent' ICR (Intelligent Character Recognition) functionality of Document Imaging Software

Prepared by Gene V. Lorica and Jeffrey Ocaya, 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 any 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) functionality of the DIS.

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

   (a)   Data capture is much faster;
   (b)   Data capture would be more faithful and accurate (*as claimed by Document Imaging Software vendors*);
   (c)   The system requires fewer resources although scanners 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.

4.   But not all is bad for the document imaging data capture technique. In fact, the accuracy of automatic extraction of entries in marked fields using optical mark recognition (OMR) is very high with less than 1% error found. 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.

5.   To exploit these advantageous characteristics of document imaging, a hybrid survey/census data capture technique was contemplated by the PNSO. Our hybrid system is a combination of Document Imaging Software and data entry application with some special traits. Characteristically, our hybrid system covers the following phases:

   (a)   *Scan the forms;*
   (b)   *Extract OMR data from the scan image of the forms using Document Imaging software;*
   (c)   *And Key-in, Key-verify, and Clean handwritten data using a program that enables encoding data direct from the scan image file*

6.   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

7.   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)   We used Document Imaging Software to extract write-in entries using Intelligent Character Recognition (ICR) and Optical Mark Recognition fields (OMR) from scan image files;
   (e)   We used questionnaires with 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.

8.   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)   We had to implement data entry verification using Centry of IMPS of 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.

B.   2002 Census of Agriculture

9.   General description of the system used:

(a) Used the same scanners (with MVCS) and Document Imaging Software but not the flawed outsourced system;

(b) Some intermittent 'bugs' of the Document Imaging Software were fixed by changing its default behaviour thru VBA programming – a functionality of the software;

(c) Used stand-alone PCs instead of a networked set of PCs as a work-around to be able to process the forms by folio/batch;

(d) To solve the not so 'intelligent' Intelligent Character Recognition (ICR) functionality of the software, all write-in data items were configured for 100% verification, that is, prompt the operator to type the entry right after the program has interpreted the data - in effect, those data items were encoded or keyed-in using the software;

(e) Final output data files are in Integrated Microcomputer Processing System (IMPS) format.

10. Limitations:

(a) Key-verification of encoded data could not be done;

(b) Data cleaning has to be done using Centry of IMPS and referring back to documents when resolving data inconsistencies as they were found;

(c) After the data capture phase, the scan image files were set-aside or archived, hence, they became of no use.

C.   2007 Mid-Decade Census of Population

11.   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); 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.

(c) Developed 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;

(d) 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;

(e) Final output batch files are in Census and Survey Processing System (CS Pro) file format.

12.   Limitations:

(a) The program had no capability for the user to reflect or 'write' corrections in the scan image file.

(b) Reusing the scan image files required reprogramming the Key-from-Image program.

D.   2010 Census of Population and Housing

13.   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 adopt a multi-purpose schematic, so that it can be used in any survey/census and in different types of data capture situations through the use of a configurable KFI template file;

(e)   Similarly as with the 2007 Mid-Decade Census of Populoation, 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 a template file for each type of form. This made Key-from-Image program a full-fledged software system;

(g)   Final output batch files are in CS Pro file format.

E.   Registry System for Basic Sector in Agriculture

14.   General description of the system used:

(a)   Used the same hardware and software in converting forms into electronic format, and the same software and technique in extracting marked fields data as the 2010 Census of Population and Housing;

(b)   The first implementation of the full-fledged Key-from-Image software

F.   Decentralized Vital Statistics System (DVSS 2011)

15.   General description of the system used:

(a)   The first implementation of the Key-from-Image software for conventional data entry of forms.

III.   KEY-from-IMAGE SOFTWARE

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

16.   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)   Data encoding with sample or 100% key-verification

17.   Using the software for other survey/census needs no reprogramming of 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 when specifying the properties and behaviour of the data items during data encoding;

(c)   Skipping patterns can also be specified using the Template Editor.

18.   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

19.   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

20.   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, while 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)

21.   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]

22.   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

E.   Localization

23.   Complete support for international localization is still under development but the output of KFI can already produce Unicode Transformation Format (UTF) text.

F.   Hardware/OS Requirements

24.   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 possibly Windows 8 with some compatibility emulation mode if necessary.

IV.   CONCLUDING REMARKS

25.   The following advantages and benefits were experienced out of the hybrid data capture system that we implemented while utilizing the strengths of the Key-from-Image program:

   (a)   Data entry of handwritten entries in the forms is much faster than conventional data entry;
   (b)   Lessened data items for encoding, since OMR data items can be automatically extracted from the scan image files of the forms, with very high reliability on accuracy;
   (c)   This design allowed for the verification of keyed-in and extracted data items;
   (d)   Accuracy of jeyed in data items ranges from 99% to 100%, based on sample key-verification in the 2010 Census of Population and Housing;
   (e)   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;
   (f)   Allows corrections to be reflected in both the data file and the scan image file of the forms.

26.   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 ICR 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.