SPC/HOPS 2010/Information Paper 7

6 July 2010

ORIGINAL: ENGLISH

SECRETARIAT OF THE PACIFIC COMMUNITY

REGIONAL MEETING OF HEADS OF PLANNING AND HEADS OF STATISTICS (HOPS)

(Noumea, New Caledonia, 12–16 July 2010)

***Statistics2020 – Developing sustainable national and regional statistical capacities***

**Session 4.4: Questionnaire processing – scanning technology**

(Document presented by the Secretariat of the Pacific Community)

**PURPOSE**

1. This paper outlines the use and comparative advantages of automated data capture for censuses and surveys in the Pacific, as well as potential applications of scanning technology for other statistical collections.
2. CSPro is still the default census and survey processing tool in the region, whether it is used as part of a traditional manual data entry approach or for data editing after automating data acquisition.

**BACKGROUND**

1. In the early 1990’s when island countries first adopted machine processing, the US Census Bureau was the main outside organization that provided technical assistance in data processing. They developed their own free public-domain software called IMPS, which later became CSPro (Census and Survey Processing System).
2. Traditionally census or survey data is collected through paper questionnaires and then manually entered into CSPRO, with most countries opting for double-entry to improve accuracy.
3. Scanning provides a viable alternative to this process, markedly improving the speed of data entry while maintaining accuracy. It also frees up resources, people and equipment, which can be deployed for other activities, and it provides much faster access to results.

**WHAT IS SCANNING?**

1. Scanning essentially involves the digitalization and interpretation of marks and text using two processes: Intelligent Character Recognition (ICR) and Optical Mark Recognition (OMR). ICR technology allows for handwritten text to be interpreted and “learned” by an application, while OMR is more straight‐forward and captures marks on a form, often a box, which has been crossed or filled in.
2. Form design and careful completion of questions are very important, as messy writing can be hard to interpret by the software. Text and mark boxes which influence how respondents write can improve error rates significantly. The series of 3 boxes below are used to enter a 3‐digit number which will be interpreted using ICR, with each number expected to be recorded in a separate box.



1. Using dropout colors (in this case green) which are invisible to the scanner, and white text boxes, the respondent is encouraged to write no more than three numbers and to evenly space them in each box. Below is what the scanner actually sees when it filters out the dropout color:



If the numbers are not evenly spaced like this, and are not written independently then they can be interpreted as another character or completely un‐interpretable.

1. Mark boxes (see examples below) can be set up in such a way that there is a threshold indicating whether or not the box is filled in, if a mistake has occurred, or if the entry is actually a speck of dirt on the form.

In the example below:

Box 1: This box has been almost completely filled in, indicating a mistake. The interpretation engine will ignore this box completely. (Boxes with more than 80% of the interior filled in are considered a mistake)

Box 2: This box has a scribble on it, or some dirt and is ignored as in 1. (Boxes with less than 20% of the interior filled in are also considered a mistake)

Box 3: This box is correctly marked and considered to be “selected” by the interpretation engine.

Box 4: This box is blank – nothing filled in, no mistakes.

**HOW DOES SCANNING WORK**

1. After reviewing several products, SDP decided to use a package called Readsoft FORMS ([www.readsoft.com](http://www.readsoft.com)). The scanning process below will be explained using this software, though the concepts are the same for other scanning application.
2. FORMS is installed in a network environment with a database like Sql Server running on a central server and workstations involved in the scanning and verification operations (see Paragraph 10).
3. FORMS comprises of a series of modules:

* 1. **Manager** – configures the questionnaire by defining each individual question and data types expected, there are also general settings, user access etc.
	2. **Scan** – controls the scanner and scans forms to TIF images
	3. **Interpret** – using rules/constraints set up for each data item on the questionnaire (through Manager module), try and interpret the data on the form
	4. **Verify** – where data has been hard to interpret or hasn’t abided by rules in the questionnaire design, present the data item to a user for modification/confirmation.
	5. **Transfer** – export data to a format (CSPRO, SAS, SPSS and others) for tabulation/processing etc
1. The *setup/scan process* works as follows:
	1. Install a copy of Readsoft Forms on the server along with SQL server (database application). This server stores all the data, the images being scanned etc.
	2. Attach a computer with the MANAGER module where we set up the system and define each question on the questionnaire ‐ this is done by physically dragging a box around the question to tell the INTERPRET module where to expect some information to interpret. We also define the type of characters to accept, such as letters for someone's name, age ranges etc (can do some pretty good error checking here!)
	3. We attach at least one computer with a scanner to the network which is running the SCAN module. The scan module controls the scanner and sends scanned images to the database on the server. Below is an image of the Kodak 1400 scanners SPC is currently using

(<http://graphics.kodak.com/docimaging/US/en/Products/Document_Scanners/index.htm>).



* 1. We attach at least one computer which is running the INTERPRET software, which then goes through and tries to "understand" all the questionnaires which were scanned. (According to the rules we set up in the MANAGER module). Any errors or uncertainties are stored in the database so the VERIFY module can check them.
	2. We open a VERIFY module. The verify module brings up each questionnaire where there is a problem and shows the verifier what it is ‐ they then change the result to what it should be. Each error is systematically displayed to the verifier to be corrected or accepted (accepted in the case it is not an error, but an exception to a rule).

*“Mass verify”* is a useful function in FORMS which is shown in operation in the image below. Letters and numbers are displayed on rows according to how they were interpreted by the software. In this case all the 1’s, then 2’s, then 4’s and 8’s are displayed. If any of them have been misinterpreted – like a number 5 appearing in the line containing 8s, the verifier selects the number “5”, and then he/she is presented with the question requiring correction on the questionnaire image. This function speeds verification up significantly.

**

* 1. The last step is to then TRANSFER the data to an output file for processing. The output file is formatted according to a transfer description (similar to a CSPro data dictionary). This file can then be processed/cleaned/tabulated in CSPro or a range of other applications such as, for example, Access, SPSS, SAC, STATA.

**other uses of SCANNING**

1. Apart from censuses and surveys, there is a wide range of applications scanning technology could be implemented to improve both speed and quality of data entry, thus contributing to getting quality data to users much faster, such as departure and arrival cards, any type of surveys, multiple choice exams, tax forms, education and health data bases.



**STANDARDISED PROCESSING TOOL**

1. Prior to the 2010 round of censuses, the idea of standardizing processing software was discussed in order to allow efficient delivery of assistance and training in the region. Many Stats offices were familiar and comfortable with IMPS and CSPro, which has a range of advantages:
	1. It was specifically designed to process census and survey data
	2. Capture, editing and tabulation of data is streamlined
	3. Freely downloadable from the internet
	4. Data can be exported to a range of other software products for further analysis i.e. SAS, SPSS, Stata, Excel.
2. With the introduction of automated data entry, there is still the need to clean and tabulate data – CSPro will always play a key role in the data capture and processing cycle.

**EDITING SCANNED DATA**

1. Although the amount of time that is normally spent on manual data entry is reduced dramatically by scanning, data editing can take much longer. With traditional manual data entry, the data is verified by double entry, which can fix 100% of keying error if full verification is engaged. Range checks, control skips and other basic checks are also performed during data capture, while in-depth checks are left for batch editing.
2. Many of the validation checks can be performed during scanning/verification, though there needs to be a balance between fast processing and preventing certain errors though, which blocks processing until an error is fixed. This is to some degree trial-and-error and reinforces the importance of a *pilot test* as part of planning for a census.
3. SPC along with several countries in the region who have already used this technology, and other partners, are trying to determine the most efficient balance between speed and quality of automated data entry.
4. Some things to keep in mind during automated entry:
	1. Scanning does not automatically fix everything. It requires more effective quality control in the field and the census office. All forms need to be checked, ordered, counted and recorded as part of the registration/control process. This then ensures that there is a good check against the number of households/forms being scanned, and EAs are correctly ordered to ensure smooth processing.
	2. Users need to keep in mind that whatever is written on the forms will go directly into the database; therefore all forms must be thoroughly checked. This is especially the case for ID items, to make sure that all forms batched together belong to the same location (e.g. island, EA) before the forms are scanned.
	3. Forms must be sorted in proper order of households and order of forms. For households which have more than 1 form, ensure that continuous forms are not mixed up within households.
	4. Serious errors can be generated if recognition rate is very low, because answer boxes are not clearly marked, or open answers are not clearly written. Recent experience has shown that the use of pens rather than pencils can help minimize this problem.

**DISCUSSION POINTS**

1. The meeting is invited to comment on:
2. The usefulness of scanning for census and survey activities, as well as other activities within NSO’s.
3. The idea of NSO’s being the “processing” centers for other departments to use – immigration cards etc

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_