UNFPA-SPC/Census Review 2012/Working paper 5
19 May 2012

ORIGINAL: ENGLISH

SECRETARIAT OF THE PACIFIC COMMUNITY
STATISTICS FOR DEVELOPMENT PROGRAMME

UNFPA-SPC REGIONAL WORKSHOP REVIEWING THE 2010 ROUND OF POPULATION AND HOUSING CENSUSES IN THE PACIFIC
(Noumea, New Caledonia, 21–25 May 2012)

Experience from the Pacific – Lessons learned for 2020

Session 6 : SCANNING, DATA ENTRY

(Document presented by Leilua Taulealo, Population Data Officer, and Phil Bright, GIS Specialist (Census and Survey), Secretariat of the Pacific Community)

1. INTRODUCTION

This paper discusses preparation for data capture (entry), whether it be manual or automated. The filling-in of forms is almost the same and the tabulation at the end of data entry is identical. Between these two processes there are various things to consider depending on whether KDE (keyed data entry) or scanning is used.

This is the first census round in which scanning has been used in the Pacific, first in Fiji (2009), then Vanuatu and the Solomon Islands (2009), and more recently in Samoa and the Cook Islands (2011). There have been a few teething issues and lessons to take into account for the 2020 census round. If scanning is correctly prepared for, and well managed, it can significantly speed up data processing and improve data quality by avoiding manual data entry errors.

Having said this, at the end of the day, the single biggest impact on data quality remains the quality of field enumeration: if bad quality data are collected, then bad quality data are captured and no amount of data cleaning and imputation will be able to produce a census dataset which is anything but questionable. Working paper 7 on managing field enumeration will address this in more detail.

2. BACKGROUND

Technology is changing, as demonstrated at the Census Technical Meeting held in Noumea in 2008. Optical Mark Recognition (OMR) and Optical Character Recognition (OCR) allow us to automate a large amount of data entry (particularly on clean, well structured questionnaires) and only require user involvement to confirm certain important pieces of data (such as ID fields) and fix errors detected through validation rules.

Scanned questionnaires can easily be linked to a database allowing efficient retrieval of forms when errors are detected, and these forms can be effectively archived for future reference. Scanning also supports on-screen coding.

Scanning isn't the answer for all countries and every application though. Many Pacific countries are relatively small and Statistics offices comparatively small as well. The initial investment in scanning technology is high for small countries, though can be repaid quickly through time savings and reuse in other activities. There is also a much more significant requirement to have good technical skills within the Statistics office – to manage the network used for scanning and also to program and manage validation checks.

3. IN-HOUSE EXPERTISE vs RELIANCE ON OUTSIDE ASSISTANCE

There is often the expectation with Census activities that outside assistance will be provided by SPC, UNFPA and other agencies. This is understandable, particularly in countries lacking staff numbers and expertise. What is critical though, is that countries take ownership of the census projects, particularly the data capture component. There is a technical skills requirement which needs to be met, or there will be too much dependence on outside assistance.

RECOMMENDATIONS FOR IMPROVEMENTS

- *When deciding between KDE and Scanning, countries should think carefully of whether or not they will have the technical expertise in-house.*
- *A successful delivery of specialized technical assistance during the recent census round has proved to be south-south type technical collaboration between Pacific NSOs, as previewed in the* Ten Year Pacific Statistic Strategy 2011-2020*, with the Vanuatu National Statistics Office sharing their scanning/data entry specialist with the Solomon Islands, Samoa and recently the Cook Island NSOs. We intend to utilize such regional technical assistance delivery, across different census and survey operational components, more frequently in future census and household survey operations.*

4. QUESTIONNAIRE DESIGN, EDITS SPECIFICATION, TABULATION PLANS, DATA DICTIONARIES

Census preparation is often left to the last minute, being *reactive* rather than *proactive*. This leads to many of the key, yet seemingly insignificant aspects of a census being rushed, or postponed.

A builder cannot successfully build a good house without detailed plans from an architect. A census project is no different. Unless countries know *why* certain questions are being asked and have a plan for *how* these questions will be asked, the data collected and disseminated, then there is a good chance that these may not be the right questions, and the results may not be what was expected. A bad plan often also means that unexpected errors creep in along the way.

A well designed questionnaire, clear edit specifications and tabulation plans, and a complete data dictionary are essential requirements for a successful data processing application. Questionnaires are looking better and better, particularly with the Pacific Model questionnaire being used as a reference by many countries. That is often where the planning stops, with edit specifications, tabulation plans and data dictionaries being compiled just before or even during data capture. The whole process is iterative and through the development of each of these components we often find that the resulting tables will not be structured correctly so we have to go back and modify the questionnaire. Subject matter specialists must be involved.

Using a model questionnaire not only means questions are being asked according to standard conventions, with correct filter questions, but also means that existing edit specifications, tabulation plans and data dictionaries do not need to be created from scratch. Countries also need to keep in mind that this means these components have been tested!

Even with all of these elements being prepared well in advance, they need to be tested, ideally through a pilot census. Everything needs to be carried out as if it were the main census, including tabulation and basic analysis to determine if the results are those being sought after. If not, appropriate, yet controlled changes need to be made.

Whether using KDE or Scanning, all the above elements are required. The only difference with scanning is that the questionnaire needs to be structured and printed correctly in order for it to be interpretable.

RECOMMENDATIONS FOR IMPROVEMENTS

- *Prepare questionnaires, edit specifications, tabulation plans and data dictionaries well in advance*
- *Subject matter specialists need to be involved in the creation of the edit specifications and tabulation plans. All validation and edit rules developed must be based on these specs.*
- *Prepare all of the above in unison so that iterative changes can be made*
- *If scanning, ensure questionnaires are correctly designed with pure drop-out colours, correctly sized mark boxes and clear recognition marks.*
- *Pencil should not be used to fill in scan questionnaires – only black pen.*

5. DOUBLE-CHECKS, ORDER OF FORMS BEFORE PROCESSING

As mentioned previously, fieldwork is the single biggest determinant of a successful census. Even if the supervision and form checking isn't ideal in the field, office controls are mandatory. It can be too time consuming to check all questions on forms in the office, but at a minimum, all ID fields need to be checked and forms ordered. If individual person forms are used, then even more careful checking needs to be done to ensure that the sum of all person forms matches the population total on the household form.

A good household listing is one of the best documents which can be used to double check form and population counts. Any mismatches are then studied more closely to figure out where the problem is.

NOTE: After data is captured and cleaned, households can be linked up with GPS points if they have been collected as a great check of completeness. Any GPS points with missing questionnaire data either were not enumerated or the questionnaire was not processed. Conversely, any questionnaires without GPS points were either not listed...or could even have been processed with the wrong Enumeration Area, Village etc.

RECOMMENDATIONS FOR IMPROVEMENTS

- *Double-check ID fields in the office*
- *Order forms prior to data processing*
- *As a post-processing exercise, link households up with GPS data if it exists to find out if forms may need to be located and processed if they have been missed.*

6. MANAGING DATA ENTRY

Data processing and tabulation will be discussed in session 8. This section will mention some of the key points to remember if scanning data prior to exporting to CSPRO.

Assuming that all questionnaires have been ordered prior to scanning, the key is to make sure that the forms scan correctly with none of them being stuck together. Manual counts should be performed and recorded on a cover page which is then verified with the automatic tally in the scanning software. If counts don't match then checks are performed.

Almost all of the CSPRO edits can be built into the scanning solution, to control the interpretation and verification of the data as it passes through the software. Any data which fails range or validation checks will then be verified manually.

Throughout the scanning process, regular backups should be performed – both of the scan database and also the scanned images. These backups can also be restored on a supervisor's machine to perform test transfers on a regular basis.

An advantage of having digital copies of the questionnaires is that they can be retrieved when checking errors in the data. These images are used in the scanning software when verifiers are performing checks and can also be used in another application like Microsoft Access. Access was used by all countries using scanning to prepare the data for CSPRO. Person and Household datasets were imported and additional validation checks performed, as well as any additional coding which was required. A little program was written in Access which then exported the data to a CSPRO dat file.

Whether scanning or using KDE, all CSPRO modules are still used. The batch editing and tabulation is performed to double check and clean the scanned data and also perform any necessary imputations.

RECOMMENDATIONS FOR IMPROVEMENTS

- *Pay closer attention to evaluation of scanning/verifying in real time.*
- *Verifiers sometimes gave the impression that they could simply push buttons without thinking – the software assists the verifier to determine the problem but doesn't always give the answer. Verifiers need stop and think.*

7. CONCLUSION

There are now multiple methods for capturing data. Countries' decision which process or technology to adopt will depend on the complexity of their questionnaires, existing in-house skills, and desired results. Even with a chosen data capture solution, poor fieldwork, resulting in messy, incomplete forms contributes to many of the problems encountered during data processing later on. Maybe we need to look at new technologies to manage this too?

One big lesson learnt over the last few years is that scanning is not magic, and cannot fix everything. It speeds up data entry, but if not managed correctly can introduce many other complexities.

There are also other technologies which were not used in the 2010 round of censuses such as PDAs (Tuvalu is planning to test them in November). With PDA unit costs dropping, this technology could become a viable solution in the next census round, allowing data validation to take place at the point of data entry (in the field), and thus adding immensely to the quality of census data.