DATA COLLECTION

CENSUS TRENDS

With the round of 2010 censuses fast approaching, many countries are preparing for one of the largest peace-time logistical exercises they will have ever undertaken

ACCORDING TO THE UNITED NATIONS Statistics Division, over 100 countries will be carrying out their national census in the two years 2010 to 2011.

The process of collecting data from every household is a major task for National Statistic Offices (NSOs) and the accurate recording and processing of that data is critical if the resultant census data sets are to be of any value. Many countries have adopted a face-to-face interview approach to collecting data in the field, using staff called enumerators. Enumerators visit households in turn and collect and record the relevant census data. A typical enumerator may be allocated to visit 500 households within a pre-defined area. The defining of these enumeration areas in itself is a major task and will use a combination of physical maps, aerial photography and satellite photography along with Geographic Information Systems (GIS) software.

VALIDATING DATA

NSOs are under increasing pressure, not only to collect the data in the field, but also to validate and disseminate accurate data in a timely fashion. NSOs are therefore actively seeking technical solutions to help them with this process. Whilst it is generally accepted that a single technical solution will not fulfil every aspect of a national census, it may speed up the process and reduce the costs associated with the majority of tasks involved.

Manual data entry, Optical Mark Recognition (OMR – the capture of tick box type data, similar to that used on multiple choice examination and lottery forms), Intelligent Mark Recognition (ICR – the capture and

> “The process of collecting data from every household is a major task for National Statistic Offices (NSOs) and the accurate recording and processing of that data is critical if the resultant census data sets are to be of any value”

interpretation of hand written text or numbers), Personal Digital Assistants (PDAs – direct data entry), pocket PCs and multi-channel are all methods currently being considered for deployment in census data capture operations.

Multi-channel refers to multiple ways of capturing the census data. For example, a statistics office might publish an online form on a website and simultaneously set up an automated telephone system for people to call and enter their data through options on a telephone keypad.

PDAs and Pocket PCs are used to collect data in an electronic form by the field worker entering data directly into the device via a screen. Data is held in the device's memory for later extraction once the unit is returned to the office or transmitted from the device directly after being submitted. They also may have Global Positioning System (GPS), camera and cellular telephone technology incorporated.

HISTORICAL VIEW

Recent research carried out on 51 countries discovered that in the last round 30 countries chose manual data entry, 11 chose ICR, seven chose OMR, two chose PDAs and one chose a multi-channel solution for their census data capture.

Of the countries that chose manual data entry, Microsoft Windows® based IMPS and CS-Pro data entry software were the most popular software applications used. Most likely because this software is provided to NSOs free of charge by the US Census Bureau.

Paper-based scanning of the census forms accounted for approximately 36 per cent of the data processing exercise using a combination of OMR or ICR technologies.

Interestingly, the research found that how smoothly a census exercise ran was less attributable to the method chosen and more attributable to how well the method was implemented and managed.

FUTURE VIEW

Indications suggest a significant shift away from manual data collection for future rounds: From the 51 countries, 15 are proposing to use ICR, 10 to use OMR, six plan to use PDAs, four intend to use manual entry and two propose to use a multi-channel solution for their next census data capture. The remaining 14 countries are as yet undecided.

As well as a migration away from manual data entry, there are regional influences on how census data is captured and processed. An example of this is the anticipated use of PDA technology in the South Pacific, The Middle East and South America. Small countries may form regional groups, sharing the same methods, experience and technology. They may also feel an element of peer-pressure from neighbouring countries to adopt the similar technologies planning to be used.

For the 2020 round of censuses, it is likely that PDA technology will become the most common method used for field based data collection due to the development of technical infrastructures, more readily available technology and reduced costs of the devices themselves. For the 2010 round of censuses this technology is likely to be trialled extensively.

Those that have yet to decide which method to deploy for their next census exercise cite uncertainties around budget and incomplete tests and trials of the technical solutions as reasons.

Those countries that have chosen are mixed in the methods they have selected. OMR and ICR feature strongly and this means that paper-based data capture will play a significant part. Paper is still a popular way of collecting data – despite the large numbers of forms to be completed; the costs are still less than any current hand-held digital solution.

Some regions are looking to implement hand-held solutions with a 'share scheme' to offset some of the capital outlay of such large numbers of digital devices with other projects. It will be interesting to see how this works in practice.

There are also technical limitations to be considered. With ICR, for example, the language used may dictate the level of accuracy that can be achieved. Specifically, the ICR recognition of some Chinese characters does not yet work well enough to be used for census data capture.

Another technology that is likely to come to the forefront of field based data capture is GPS. Significant benefits to all users of census data can be achieved if the data has a geographical reference. Software used to disseminate the census data is becoming more visual and the incorporation of maps into this dissemination process is becoming increasingly attractive. Typically data has been displayed in tabular form to users and so the use of maps to show trends and spread of the census data is an effective tool. With the technical advancements of GPS positioning accuracy and the cost of such devices falling, associating GPS information to the census data collected is likely to become common place.

With IT infrastructures becoming more readily available to the general population of many countries, more NSOs are likely to offer multi-channel methods of collecting census data during the 2020 round of censuses. The Internet will doubtless play a large part if there is to be a shift away from face-to-face data collection. Again, for 2010 it is likely that many countries will trial multi-channel methods.

Looking beyond 2020, it is likely that census data will be captured on an ongoing basis via other exercises and processes, probably through local government. As government departments within countries become more coordinated and information is shared more freely, the need for large census data collection exercises could disappear altogether.

> "Looking beyond 2020, it is likely that census data will be captured on an ongoing basis via other exercises and processes, probably through local government. As government departments within countries become more coordinated and information is shared more freely, the need for large census data collection exercises could disappear altogether"

Andy Tye is global business leader for census at DRS Data Services Limited. DRS has been involved in more than 15 national population registration projects around the world.