Australian Bureau of Statistics

statistics for informed decision making

Census Data Capture: ABS Experience 1991 to 2006

Noumea

February 2008

Optical Mark Recognition (OMR)

- 1991 and 1996
- A quick, cost effective and high quality way of capturing households and persons within households
- Most Census information can be appropriately collected via tick-boxes
- Technology has become more reliable over time

Optical Mark Recognition (OMR)

- Quality of printed forms is critical to success
- Need the right drop out colour and form designed to meet scanner requirements
- Quality assure OMR process
- Routine preventative maintenance
- Need in-house expertise (not a black box)
- Data Reformat – business rules

Imaging

- 2001 introduced imaging
- Images replaced paper forms for all post data capture processing
- Efficiency gains in Editing and Computer Assisted Coding (CAC) particularly
- Supports evaluation and validation
- Efficient reprocessing for errors

Imaging

- Developed application to capture part of image for a question (“snippets”)
- Call up form snippets relevant for editing or coding process

- OHAS issues with intensified screen-based work

ICR and Beyond

- 2001 and 2006 use ICR (IBM Intelligent Form Processing) for capture of write-in responses
- Aim is efficiency through Automatic Coding of captured text responses
- Critical to design form to suit ICR

ICR

1. Recognition engine attempts to recognise characters

2. Run (usually partially) recognised text string against Automatic Coding indexes of anticipated responses

3. If no match go to character repair (computer assisted)

4. Run repaired text against AC indexes

5. If no match go to CAC

ICR - Tactical

- Deciding the appropriate probability for accepting characters from ICR
- Deciding what is an acceptable match in AC
- Deciding when to go to character repair and when to go straight to CAC
- Understanding and optimising algorithm for AC matching
  - ABS developed a customised AC algorithm for multi-field coding (eg addresses)

ICR – Quality Assurance

- Need to understand the solution well to manage quality – develop in-house expertise
- Substantial investment required in Auto Coding Indexes
- Comprehensive testing of ICR engine, AC algorithms/indexes and outputs from current “real world” responses
- Extensive validation effort to detect and correct systematic false positives – will be very visible in output

ICR - Costs

- 2006 Census purchased 12 Inotec Scamax 510 document scanners.
- A total of 65.5 million pages were scanned over 71 working days.
- The images from the scanners were transferred to 18 form processing pcs where the IFP application processed the images using ICR technology.
- The resultant data and associated images were then loaded into Oracle and image data stores
- IBM contract and scanners cost $A2.5M

Where to for 2011

- Further refine ICR, Auto Coding algorithm and AC indexes to improve AC match rates >>>> less CAC
- More macro approach to quality assurance
- Promote eCensus response channel