



Census Paper Data Capture Quality

Decennial Response Integration System (DRIS) Paper Data Capture - Leveraging Automation,
Ensuring Quality, and Containing Cost for the 2010 Census

March 29, 2011

DRIS PMO

U.S. Department of Commerce, Bureau of the Census

Suzanne Fratino

suzanne.fratino@census.gov, 301-763-4134

Alan Berlinger

alan.j.berlinger@census.gov, 301-763-4000

Executive Summary

The U. S. Department of Commerce's 2010 Decennial Census generated a massive volume of respondent-filled Census questionnaires. The Census Bureau's Decennial Response Integrated System (DRIS) program was tasked with transforming every checkbox question and handwritten field on each paper form into a digital format that allowed for quick, accurate, and efficient tabulation.

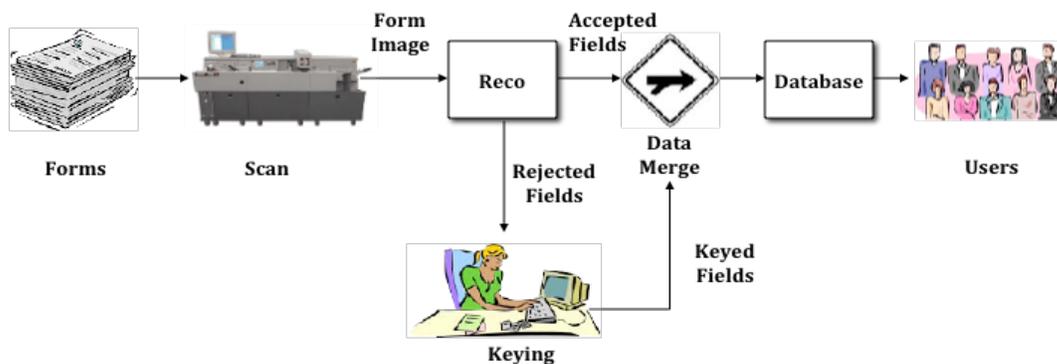
The DRIS program developed and deployed innovative and highly automated technologies that resulted in exceeding data quality standards, vastly increased productivity, and significantly reduced cost. The use of automation enabled the real and near real time identification of small pockets of error during production. These errors were corrected as part of the normal quality processes, as defined in the DRIS Data Quality Management Plan (DQMP) which specified Service Level Agreements on data quality outcomes.

The DRIS program accurately captured complex data for over 165 million census forms employing high-speed color image processing and data capture technology. They developed and applied numerous tools to measure and ensure data quality, and achieved a very high accuracy rate. The DRIS team performed this task within a short six month timeframe, and finished operations under budget.

Challenge

Under constitutionally mandated deadlines, the DRIS program was tasked with capturing data from millions of forms from 69 different form types resulting in 555,000,000 8 ½ x 11 sheets of paper. At the beginning of the Census, DRIS received a massive wave of returned forms, checking in close to 8 million forms a day, with an ultimate requirement for all data captured and delivered within six months. DRIS was challenged to meet 99.8% accuracy for automated capture of checkbox questions (OMR or Optical Mark Recognition), 99% accuracy for automated capture of handwritten fields (OCR or Optical Character Recognition), and 97% accuracy for keyed fields.

The following diagram depicts the paper data capture process at a high level (the box labeled 'Reco' represents the automated mark and character recognition, OMR and OCR):



Solution

The DRIS paper data capture process was open and transparent allowing for multiple views of data to be presented to analysts as part of a CMMI level 5 best practice of continuous improvement. For the first time, Census simultaneously ran production operations and performed data quality evaluations in parallel. These evaluations previously took months to years to complete and took place only after production ended. The DRIS technology and automation advances allowed the Census Bureau to apply evaluation metrics to production quality outcomes in real time.

In addition to maximizing the automation related to paper data capture, the DRIS solution leveraged automation of data quality tools, integrating both to enable the continuous improvement.

The DRIS program developed major advances in technology, test, and automation by designing and deploying innovative systems and tools that rapidly identified capture and respondent errors many times faster than the traditional, manually intensive human intervention methods. For example, they developed the automated Production Data Quality (PDQ) system, which

validated and verified data quality in real time, resulting in a 32-fold increase in efficiency over recently-used manual techniques.

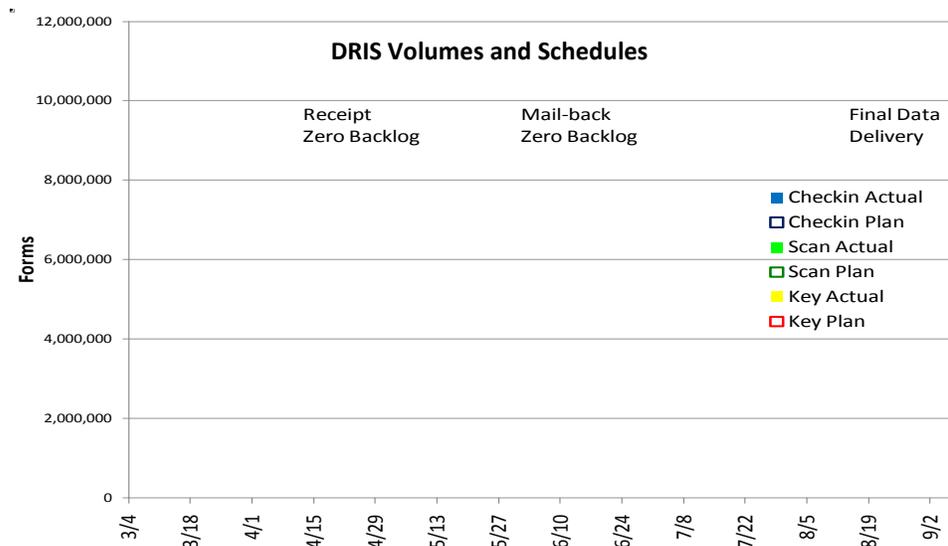
To do this, DRIS combined multiple data quality approaches built into a flexible, highly functional and scalable architecture. Examples of the data quality approaches included implementing an acceptance quality control plan to manage the quality of the work product, fit-for-use matching algorithms, integrating real-time data quality reporting systems including the use of Business Intelligence tools, implementing Association Logic analysis tools to detect patterns, developing high performance databases, and providing frequent reporting intervals to provide data to manage the operations at all levels.

Integrating the data quality solutions with the data capture system enabled not only savings in automated capture but also enhanced the quality.

Results

The DRIS paper capture system demonstrated state-of-the-art success. The Optical Character Recognition (OCR) field accuracy rate was 99.56% compared to a required 99%; the operator field keying accuracy rate was 98.61% compared to a required 97%; and the Optical Mark Recognition (OMR) accuracy rate was 99.98% compared to a required 99.8%. One significant example of the value of multiple data quality approaches was identification of overstating the multiracial populations where respondents marked multiple boxes in error or respondents marked boxes indicating a negative response vs. a positive response. The timely identification of this issue allowed the Census to revise capture rules and DRIS to reprocess the forms.

The DRIS paper data capture automated solution eliminated over 80% of the traditional manual data entry associated with a Census data capture operation, reducing cost by hundreds of millions of dollars.



Lessons Learned

- The DRIS program enabled technology, testing, and automation advances that will continue to support and inform Census technology through the decade as well as aid management and design decisions related to planning the 2020 Census.
- The application of the combined multiple data quality approaches, including an acceptance quality control plan to manage the quality of the work product, fit-for-use matching algorithms, real-time data quality reporting systems, and Association Logic analysis tools, allowed for optimization of cost and quality for automated data capture.
- The use of novel synthetic test materials and testing methods contributed to the overall success of the system.

Disclaimer

- References to the product and/or service names of the hardware and/or software products used in this case study do not constitute an endorsement of such hardware and/or software products.