Testing Production Data Capture Quality

K. Bradley Paxton, Steven P. Spiwak, Douglass Huang, and James K. McGarity

ADI, LLC
200 Canal View Boulevard, Rochester, NY 14623
brad.paxton@adillc.net, steve.spiwak@adillc.net, doug.huang@adillc.net, jim.mcgarity@adillc.net

Executive Summary

Testing the data capture quality from a production forms processing system's final output is a difficult and costly process if done with manual data entry methods. To overcome this problem for the U.S. Census Bureau's 2010 Decennial Response Integration System (DRIS), we created Production Data Quality (PDQ), a specialized data capture system whose main purpose is to independently and efficiently assess the data quality of a production data capture system's final output. It accomplishes this by semi-automatically producing a high quality Truth* for the sampled production final output data, and then using that Truth to score the final production results and identify errors. During Census 2010, PDQ was used to independently score and track DRIS error and reject rates and to provide near-real-time analysis in order to uncover pockets of error in DRIS production output. Production spanned from March 4, 2010 through September 7, 2010; and, during that time, 865,000 forms representing more than 50 distinct form types were processed by PDQ, the Truth obtained, and the final DRIS data output scored.

We found that DRIS performed well within required service level agreements (SLAs), with respect to its overall accuracy rate performance in Optical Character Recognition (OCR), write-in keying and check-box recognition. These SLAs were stringent, requiring 99% field-level accuracy for OCR, 97% for write-in keying and 99.8% for check-box questions. Furthermore, a comparison to Census 2000, after rescoring DRIS according to the legacy soft-match scoring criteria employed in 2000, showed that DRIS slightly exceeded Census 2000 total write-in accuracy while improving production efficiency by 25%.

However, many data quality issues, clusters of error, and causes of unexpected response patterns are sometimes masked by overall error rate metrics but were nevertheless uncovered by PDQ, including an overstatement of multiple mark check-box responses on respondent-filled forms and the impact of erasures on enumerator-filled forms. PDQ's daily feedback loop to Census during production was critical in the identification of these and other issues that were promptly addressed by DRIS. Remediation took various forms ranging from the repair of form definitions, modification of keying rules, and, in a few critical cases, reprocessing of over 3 million forms.

The PDQ concept may be extended to determine Truth of other types of production classifiers, such as record linkage systems, in addition to paper data capture systems.

---

* We always use a capital T for the "Truth," if only out of respect for how difficult it is to obtain. This Truth is sometimes called "ground truth" in the data capture business, and refers to precisely correct data representing the respondent's writing on the form. This Truth is then used to score the final production output data for accuracy.

[This paper was published in the Federal Committee on Statistical Methodology (FCSM) 2012 Conference Proceedings]

PDQ Description

PDQ inputs a sample of production source images and corresponding captured data and determines the relevant Truth with a high degree of precision, while leveraging software automation and sophisticated statistical design[1] to minimize effort and operational cost. When coupled with immediate sampling of production data, PDQ provides near-real-time data quality measurements and feedback. The PDQ system employs independent Optical Character Recognition (OCR) and Optical Mark Recognition (OMR) engines and expert human Data Capture Analysts (DCAs) in a manner that avoids systemic biases associated with internal verification processes.

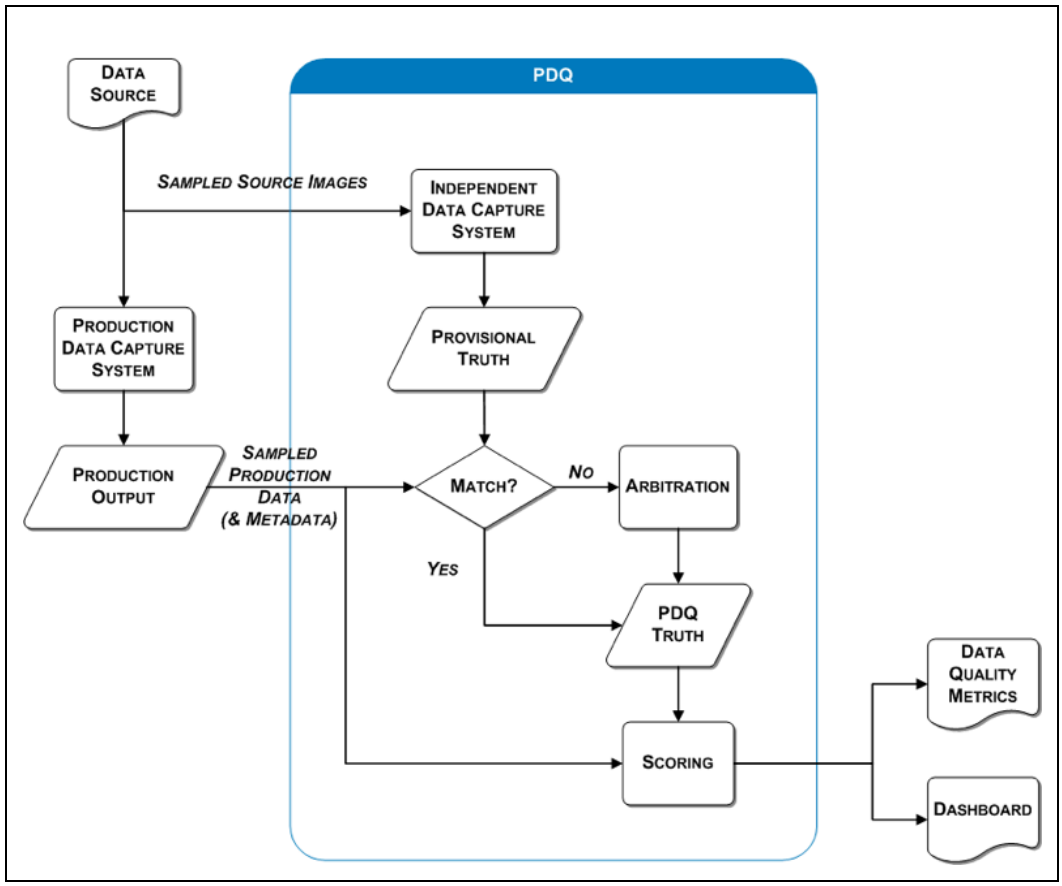


Fig. 1 – A Block Diagram of PDQ with the Production Data Interface

A randomly sampled set of images, corresponding production (field level) data and associated metadata was selected from the much bigger population of production images captured by DRIS (see Fig. 1 above). This sample of electronic files was sent to PDQ, which processed them via its "Independent Data Capture System." Write-in and check-box field data was captured using automated OCR and OMR engines, and human DCAs (using the "Form Completion" module, a part of the "Independent Data Capture System") for low confidence fields. This data, labeled as "Provisional Truth," was then compared to the corresponding DRIS production results using a hard match algorithm (i.e., the same characters, the same number of characters, in the exact same order) to determine which fields matched. PDQ was purposely implemented with a different OCR and OMR engine than DRIS. The vast majority of field comparisons matched; and, since different technologies were used to capture the data, there was very high confidence that the matched fields represented the correct respondent data. These matches were automatically labeled as "PDQ Truth." A small number of remaining fields were reviewed by DCAs (using "Truth Scrubber," represented in the workflow diagram above as "Arbitration") to complete the set of PDQ Truth.

Once the hard match PDQ Truth was obtained, the sampled DRIS data was scored against it using the following field matching algorithms:

Field Type	Matching Algorithm
Alpha write-in	Jaro-Winkler Match algorithm
Alphanumeric write-in	Jaro-Winkler Match algorithm
Numeric write-in	Integer Match
Check-box	Hard Match

Table 1 - Matching Algorithms used to Evaluate DRIS Data Capture Quality

PDQ was the only system used to confirm the accuracy of DRIS check-box processing both prior to and during production, because OMR is so good it is hard to test.

DRIS Scoring Results

Among the most important aspects of PDQ's responsibilities during Census 2010 data capture production was the task of scoring DRIS final output against the three SLA accuracy rate requirement categories: write-in fields captured by automated processing (OCR), write-in fields captured by manual keying (Write-in Keying), and check-box fields captured by either automated processing or keying (Check-box). Table 2 summarizes the PDQ findings showing that DRIS exceeded the minimum data quality requirements set by the Census Bureau in each category, while Table 3 shows how DRIS performed relative to its accept rate design goals.

Requirement Name	Field Type	Capture Methods	DRIS Accuracy Rate Requirements	Weighted Accuracy Rate	Conclusion on DRIS Performance
OCR Accuracy	Write-in	OCR	≥99.0%	99.56%	Exceeded Min. Requirement
Write-in Keying Accuracy	Write-in	KFI, KFFI	≥97.0%	98.61%	Exceeded Min. Requirement
Check-box Accuracy	Check-box	OMR, KFI, KFFI	≥99.8%	99.98%	Exceeded Min. Requirement

Table 2 - Overall DRIS Accuracy Rates Compared to DRIS Requirements
Note: In this table, KFI means Key From Image, and KFFI means Key From Full Image**.**

Design Goal Name	Field Type	Capture Methods	DRIS Accept Rate Design Goals	Weighted Accept Rate	Conclusion on DRIS Performance
OCR Accept Rate	Write-in	OCR	≥80.0%	86.44%	Exceeded Design Goal
OMR Accept Rate	Check-box	OMR	≥99.0%	98.08%	Approached Design Goal

Table 3 - Overall DRIS Accept Rates Compared to DRIS Design Goals

Data Analysis Using the Data Dashboard

The PDQ Dashboard provides an intuitive user interface, allowing the analyst to "drill down" into increasing levels of detail, including processing history as well as the original source image via an Image Viewer. The PDQ Dashboard's unique color-coding scheme focuses the analyst's attention on pockets of error occurring within the data, even for data meeting SLA requirements at the aggregate level.

Like PDQ's scoring function, the PDQ Dashboard has a flexible scoring implementation, enabling the analyst to select from multiple matching algorithms via a configuration option. The PDQ Dashboard supports advanced root cause analysis through built-in configuration, filtering, sorting, and *ad hoc* query capabilities. It also enables further analysis using popular off-the-shelf software packages via data export, as well as integration with direct database queries performed outside the Dashboard. The PDQ Dashboard provides management summary information, including SLA verification, at the aggregate level as well as weighted aggregate results by sampling rate(s). It can also show more detailed engineering level information, an example of which is shown below.

A tangible (but fictitious) example is presented in the next four figures. In Fig. 2 below is shown an engineering level dashboard screen for OCR results.

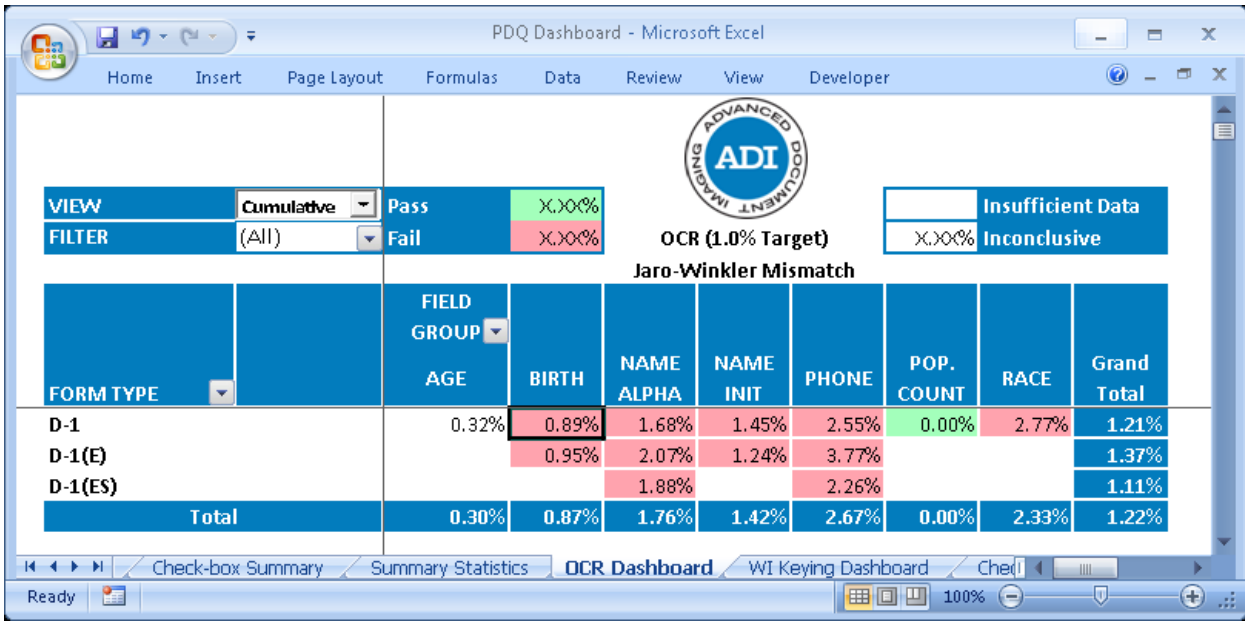
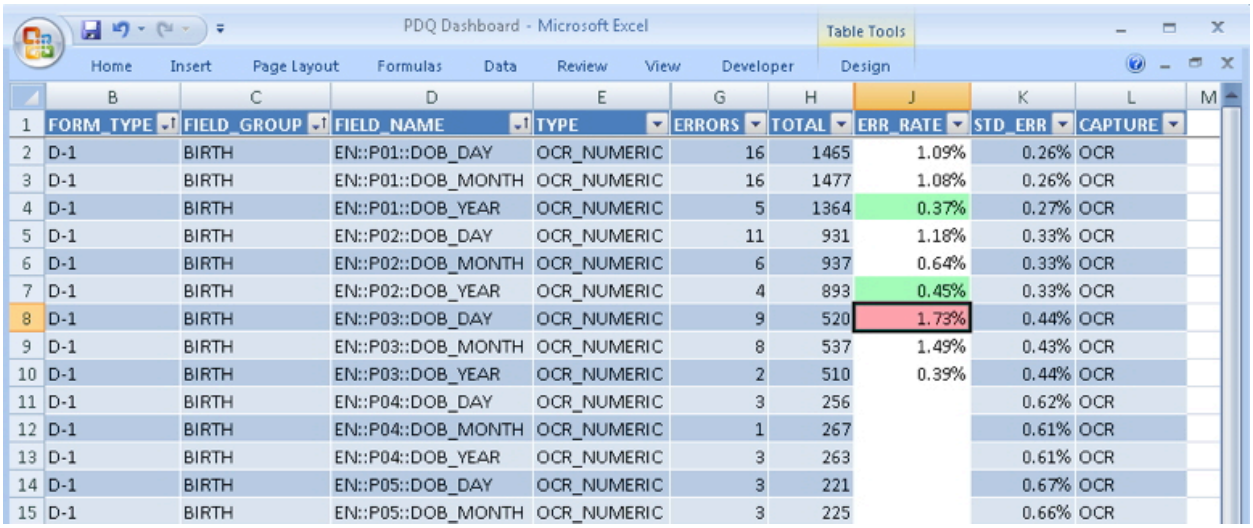


Fig. 2 – Top Dashboard Screen for OCR Results

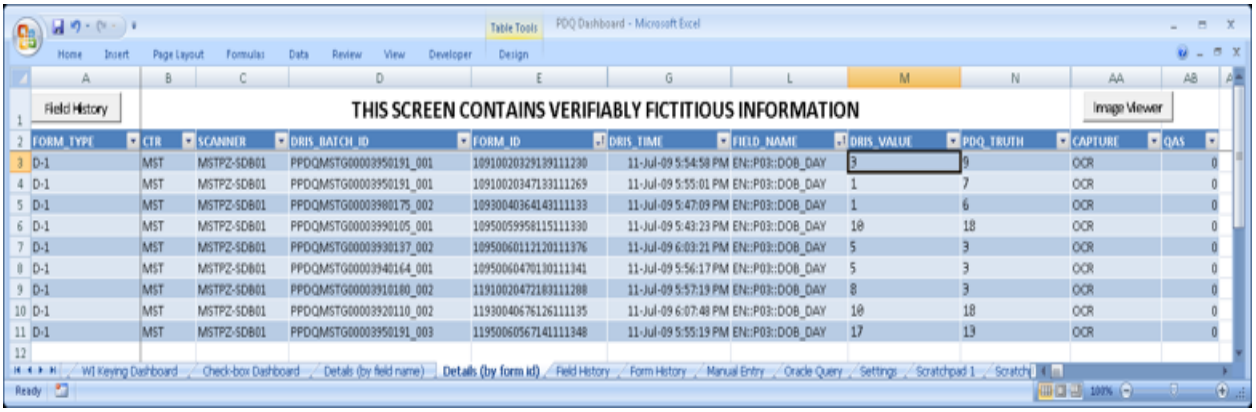
These results show field-level error rate data by form type (rows) and by field type (columns) for a particular (Jaro-Winkler) soft match. Only three form types are shown here for clarity, but there were more than 50 form types sampled in DRIS production. The target (max) error rate is 1%, and for the aggregate results, the rectangles were most always green. However, in this engineering view, a rectangle will be red if there is a pocket of error inside of the data somewhere, which is useful for analysis. The rectangle outlined in black shows 0.89% error rate for D-1 BIRTH fields, so that means there is a pocket of error inside that particular group of sampled fields. If we drill down to the next level, we get the screen shown in Fig. 3 below.



	B	C	D	E	G	H	J	K	L	M
1	FORM_TYPE	FIELD_GROUP	FIELD_NAME	TYPE	ERRORS	TOTAL	ERR_RATE	STD_ERR	CAPTURE	
2	D-1	BIRTH	EN::P01::DOB_DAY	OCR_NUMERIC	16	1465	1.09%	0.26%	OCR	
3	D-1	BIRTH	EN::P01::DOB_MONTH	OCR_NUMERIC	16	1477	1.08%	0.26%	OCR	
4	D-1	BIRTH	EN::P01::DOB_YEAR	OCR_NUMERIC	5	1364	0.37%	0.27%	OCR	
5	D-1	BIRTH	EN::P02::DOB_DAY	OCR_NUMERIC	11	931	1.18%	0.33%	OCR	
6	D-1	BIRTH	EN::P02::DOB_MONTH	OCR_NUMERIC	6	937	0.64%	0.33%	OCR	
7	D-1	BIRTH	EN::P02::DOB_YEAR	OCR_NUMERIC	4	893	0.45%	0.33%	OCR	
8	D-1	BIRTH	EN::P03::DOB_DAY	OCR_NUMERIC	9	520	1.73%	0.44%	OCR	
9	D-1	BIRTH	EN::P03::DOB_MONTH	OCR_NUMERIC	8	537	1.49%	0.43%	OCR	
10	D-1	BIRTH	EN::P03::DOB_YEAR	OCR_NUMERIC	2	510	0.39%	0.44%	OCR	
11	D-1	BIRTH	EN::P04::DOB_DAY	OCR_NUMERIC	3	256		0.62%	OCR	
12	D-1	BIRTH	EN::P04::DOB_MONTH	OCR_NUMERIC	1	267		0.61%	OCR	
13	D-1	BIRTH	EN::P04::DOB_YEAR	OCR_NUMERIC	3	263		0.61%	OCR	
14	D-1	BIRTH	EN::P05::DOB_DAY	OCR_NUMERIC	3	221		0.67%	OCR	
15	D-1	BIRTH	EN::P05::DOB_MONTH	OCR_NUMERIC	3	225		0.66%	OCR	

Fig. 3 – Level 2 Dashboard Screen for OCR Results

This screen actually goes down to Person 12 (P12), but is restricted here to P05 for legibility. Here we see a 1.73% error rate for D-1 BIRTH DOB_DAY, (which means the numeric day of birth for person 3). The screen shows there were 9 errors out of 520 fields sampled. If we wish to find out why this is, we drill down again and get the screen of Fig. 4 below.



	A	B	C	D	E	G	L	M	N	AA	AB
2	FORM_TYPE	CTR	SCANNER	DRIS_BATCH_ID	FORM_ID	DRIS_TIME	FIELD_NAME	DRIS_VALUE	PDQ_TRUTH	CAPTURE	QAS
3	D-1	MST	MSTPZ-SDB01	PPDQMSTG00003950191_001	10910020329139111230	11-Jul-09 5:54:58 PM	EN::P03::DOB_DAY	3	9	OCR	0
4	D-1	MST	MSTPZ-SDB01	PPDQMSTG00003950191_001	10910020347133111269	11-Jul-09 5:55:01 PM	EN::P03::DOB_DAY	1	7	OCR	0
5	D-1	MST	MSTPZ-SDB01	PPDQMSTG00003980175_002	10930040364143111133	11-Jul-09 5:47:09 PM	EN::P03::DOB_DAY	1	6	OCR	0
6	D-1	MST	MSTPZ-SDB01	PPDQMSTG00003990105_001	10950059958115111330	11-Jul-09 5:43:23 PM	EN::P03::DOB_DAY	10	18	OCR	0
7	D-1	MST	MSTPZ-SDB01	PPDQMSTG00003930137_002	10950060112120111376	11-Jul-09 6:03:21 PM	EN::P03::DOB_DAY	5	3	OCR	0
8	D-1	MST	MSTPZ-SDB01	PPDQMSTG00003940164_001	10950060470130111341	11-Jul-09 5:56:17 PM	EN::P03::DOB_DAY	5	3	OCR	0
9	D-1	MST	MSTPZ-SDB01	PPDQMSTG00003910180_002	11910020472183111288	11-Jul-09 5:57:19 PM	EN::P03::DOB_DAY	8	3	OCR	0
10	D-1	MST	MSTPZ-SDB01	PPDQMSTG00003920110_002	11930040676126111135	11-Jul-09 6:07:48 PM	EN::P03::DOB_DAY	10	18	OCR	0
11	D-1	MST	MSTPZ-SDB01	PPDQMSTG00003950191_003	11950060567141111348	11-Jul-09 5:55:19 PM	EN::P03::DOB_DAY	17	13	OCR	0
12											

Fig. 4 - Level 3 Dashboard Screen for OCR Results

Here, we can see that, for example, (in row three) that DRIS read a number for (date of birth) day as a "3," whereas PDQ read it as a "9." If one is then insanely curious as to how that could happen, you can drill down one more time to the screen shown in Fig. 5 below.

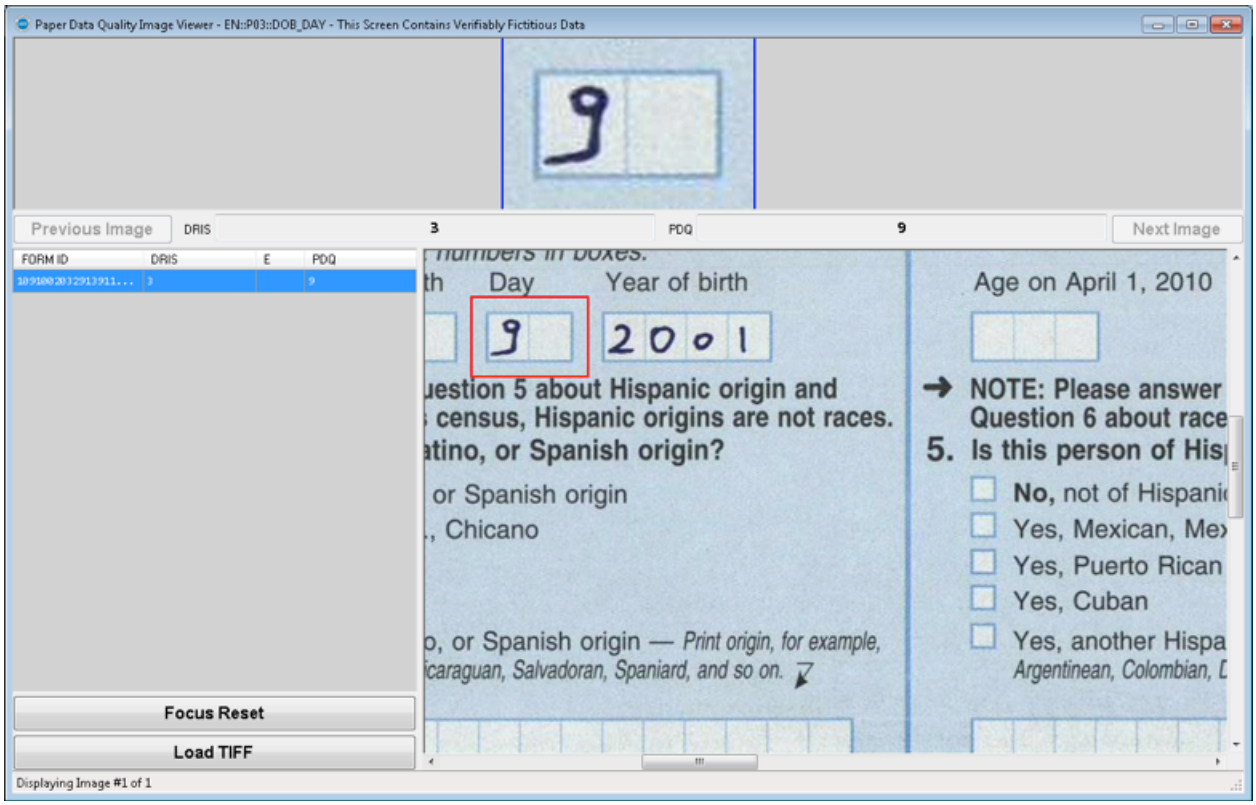


Fig. 5 - Level 4 Dashboard Screen for OCR Results

Here we now see the digit in question, which was read by the OCR as a "3," whereas PDQ claimed it was a "9." This is rather a curiously shaped number "9," to be sure, and is shaped differently than most hand-written nines. We know the engines that perform OCR do a great job, but they are trained neural-net programs that do make occasional mistakes even on apparently simple cases such as this. However, the overall numeric hard-match field error rate for the DRIS OCR system was 0.39%, and so it didn't happen very often.

Production Outcomes

PDQ operated in parallel with DRIS to provide near-real-time feedback on sampled forms. Two shifts of 10 DCAs each processed write-in and check-box fields on approximately 865,000 forms over a period of six months.

Initially, PDQ was tasked with capturing both write-in and check-box fields for selected form types only, while capturing only check-box fields for the remaining form types in the PDQ sample. However, due to PDQ's efficiency in processing check-box fields, PDQ's scope was expanded (mid-production) to include capture of both write-in and check-box fields for *all* form types in the PDQ sample. This was done at *no* incremental cost to the Census Bureau.

At the completion of production, the PDQ system with its stored results in a database remains the largest collection of "truthed" Census-related data available, and may well be the largest collection of "truthed" handwriting samples as well. In July 2011, the PDQ system moved to Census Headquarters in Suitland, Maryland, and is being maintained by the Decennial Automation Contracts Management Office (DACMO).

Important DRIS data quality issues discovered and addressed during production through PDQ analysis are described in the section below.

Sources of Error and Other Issues Found by PDQ

The unique analysis tools provided by PDQ led to the discovery of the following issues during production, most of which would not have been found without PDQ's contribution to DRIS's continuous improvement cycle:

1. Overstatement of multiple mark check-box responses (particularly in the race question)

Early in the data capture process, the Census Bureau expressed interest in the error rates associated with the Race, Hispanic and Coverage fields on questionnaires completed by respondents. A particular concern was the number of multiple-check responses, and the error rate within that subgroup. In part, the close arrangement of check-boxes on the paper forms resulted in marks that inadvertently covered more than one box; DRIS OMR and check-box keying often made errors in these cases.

In March 2010, the Census Bureau asked the PDQ team to investigate. The results showed that DRIS was successfully capturing check-box fields with an error rate that met Census Bureau requirements. However, as PDQ analysts drilled down to smaller and smaller subgroups, pockets of error were revealed. For example, the error rate for D-1 race fields where DRIS reported multiple check marks before intervention was 18.66%.

The PDQ team's analysis indicated multiple causes, originating with both respondent behavior and DRIS processing. Some respondents completed check-box questions in novel ways. For example, some entered implausible data in an apparent effort to protest their offense or dissatisfaction with the question. Others used a "yes" mark style to indicate what they *are*, and a "no" style to indicate what they *are not*. Some of these response patterns were seen for the first time during the DRIS 2010 operation, possibly because PDQ provided the tools to efficiently isolate and examine such cases. DRIS context checking rules and DRIS keying rules also played a role in the elevated error rate.

DRIS decided to take corrective action for the processing of these fields. DRIS also decided to reprocess the forms received to date. The data capture quality improved dramatically after corrective action was taken and reprocessing was complete.

2. Distinguishing incomplete erasures from light responses (such as pencil)

The 2009 decision to use paper forms, rather than hand-held electronic devices, for the 2010 Non-Response Follow-Up (NRFU) operation required DRIS to capture data from tens of millions of forms completed in pencil by field enumerators. Not only did the NRFU paper forms increase the total number of forms processed at PDCCs, it also increased the importance of correctly capturing pencil marks and erasures. In Census 2000, it was virtually impossible to differentiate erasure marks on the bitonal form images. For Census 2010, full color images were captured which contained far more information, enhancing the ability to capture the respondent's intent.

In response, the PDQ team monitored DRIS's handling of pencil marks and erasures, before and during the production period. Conclusions were:

- DRIS accuracy rates for enumerator form types completed in pencil exceeded the target rates for all field types.
- There was an elevated error rate among erased fields, but erased fields accounted for only about 1% of all fields. The net result was that erasure errors were roughly a quarter of all errors, and DRIS accuracy rates exceeded the target rates even with erased fields included.
- Early in the production process, the PDQ team informed the Census Bureau and the DRIS contractor of elevated error rate among erased fields. The agreed-upon response was the modification of software and keyer instructions. This change reduced the erasure error rate substantially.
- If the business rules change had not been implemented, error rates would have been significantly higher, but still within the target error rates.

A close examination of erased fields indicates that erasures on enumerator forms are not typically the result of data entry errors. Instead, they represent changes in the case history and available information as the enumerator(s) makes diligent attempts to collect accurate information under difficult circumstances.

3. Misalignment of Puerto Rico form images

In March 2010, PDQ engineers noticed an unusual number of check-boxes on D-1PR(S) forms were erroneously recorded as checked when, in fact, they were really blank. They traced the problem to poor alignment of the form images with the standard "template" image. The poor alignment caused the answer boxes to appear in the "wrong" place, such that the automated OMR process misinterpreted some answer box borders as respondent marks. Once the DRIS team was alerted, corrective action was taken, and data quality improved dramatically.

4. High DRIS keying workload due to Proxy fields on Non-Response Follow Up (NRFU) forms

PDQ detected an abnormally high DRIS keying workload on enumerator forms due to Proxy fields. At Census request, DRIS discontinued reject keying of these fields, treating them as OCR only, and PDQ removed them from overall DRIS scoring.

5. Keying of ambiguous responses

Early in production, PDQ discovered that many DRIS keyers were omitting ambiguous characters, keying a blank instead of the difficult character. The DRIS team responded by reinforcing the keying rules training which instructs keyers to "key what they see" to the best of their abilities.

6. Middle initial on Spanish language forms

PDQ detected multiple letter responses in the middle initial fields on Spanish language forms, indicating the translated label for the field could be construed to mean "initial here" rather than "middle initial." The DRIS team, working with Census stakeholders, responded by reprogramming DRIS to heuristically select the most likely middle initial for inclusion in the final Census data set.

7. Alpha characters in Map Spot fields

PDQ discovered that alpha characters were present in the Map Spot field, which is defined as a numeric field. DRIS keyers either blanked the alpha characters or substituted them with numbers, causing a high error rate for this field. PDQ subsequently removed this field from scoring at Census request.

8. Refusal response recognition ("N/A" v. "NIA")

PDQ discovered that DRIS OCR was interpreting fields with refusal responses (for example, "N/A") as actual responses ("NIA"). PDQ DCAs correctly interpreted these fields as blank, revealing pockets of errors in the affected fields. The DRIS team responded by automatically rejecting such refusal responses to the keying process, where a DRIS keyer would have the opportunity to properly interpret them.

Keying Efficiency and Truth Precision

The PDQ system provides an efficient and precise method for determining the Truth of a data capture system's final output. Prior to the creation of PDQ (and still common today), the standard procedure for determining the Truth of production data capture operations was to manually perform "Double Key and Verify" (DK&V), which is costly, time-consuming, and prone to error.

A detailed study by Huang[2] has shown that the keying efficiency of PDQ relative to DK&V was about 28 times. In other words, for a given test sample of production forms, the quality assurance keying effort using PDQ is nearly 28 times less than for the same work performed using DK&V. This means that the same amount of test sampled forms can be processed with $1/28^{th}$ the QA keying staff, or that twice as many samples can be processed with $1/14^{th}$ the QA keying staff, etc, to suit the client's need. For example, for future applications, the relative efficiency of PDQ can be increased from 28 times up to 38 times or even more, depending on the particular application test requirements.

It is also shown in this same study that the resultant Truth error using PDQ is from 6 to 10 times less than DK&V, depending on the accuracy of the keyers. It is rather difficult to estimate the Truth precision, (or if you prefer, correctness of the standard against which the final production accuracy is determined), but an upper bound for Truth error rate for DRIS alphabetic write-in fields was given as 0.008%, which is considered extremely low for this purpose.

Conclusions

The use of PDQ in DRIS 2010 to independently determine and track error and reject rates, and to provide analysis tools that led to the near-real-time discovery of pockets of error, was very successful. Because it creates precise Truth at low cost, it will be an essential part of ensuring the data capture quality of future Decennial Census operations and other surveys. As multiple data capture modes (phone, internet, and computer assisted enumeration) become integral to the next Census, possibly supplemented by record linkage systems, opportunities for data quality issues to arise will increase, as will the likelihood of such issues being masked by the sheer complexity of the operations. The analytical Dashboard component of PDQ has the ability to lead analysts to troublesome subsets deep within the data even if the aggregate data quality is exceptional.

Future Work

We observed in attending the FCSM 2012 Conference that there is growing government agency interest in record linkage to improve data and lower costs. The PDQ system described herein is extensible to testing record linkage systems operating on production data, enabling valuable quantitative metrics to be obtained such as false positives and false negatives. This is done in part by replacing the "Independent Data Capture System" of Fig. 1 with an "Independent Record Linkage System." This new approach (RLPDQ) is in the design stage and will be the subject of a future paper.

References

1 Paxton, K. B., Spiwak, S. P., & Huang, D. (2009). *Truthing Production Data Capture*, Joint Statistical Meetings (JSM) Proceedings, Government Statistics Secti*on* (pp. 494-500). Washington, D.C.: American Statistical Association.

2 Huang, D. (2011). *Balancing Truth Error and Manual Processing in the PDQ System.* Rochester, NY: MS Computer Science Thesis, Rochester Institute of Technology.