

FEDCASIC 2009

The Agricultural Census Is Not Just A
Big Survey

David McDonell
March 18, 2009



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



FEDCASIC 2009

Objectives

- ▶ Overview of Census Program
- ▶ Failures and Successes
- ▶ Impact on Edit System
- ▶ Applying Lessons Learned on
Census to Surveys



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



FEDCASIC 2009

NASS Organizational Structure

Headquarters - (Washington, D.C.)
400 Employees

Prepare for Survey

Samples

Instruments

Training/documentation

Administer



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



Instruments

- Blaise program
- Analysis program
- Summary programs
- Estimation programs
- Publication programs

Estimation manual

Survey administration manual

Interview's manual

FEDCASIC 2009

NASS Organizational Structure – Continued

Field Offices (FO)

44 locations

600+ employees

2,000+ enumerators

Conduct Survey

Collect

Edit

Analyze

Set estimates for their state(s)



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



FEDCASIC 2009

Typical NASS Survey

1-2 weeks of data collection / editing /
analysis / summary

Results published within a few weeks of
reference date



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



Able to touch and or review every record in a survey through editing and analytical review

Can be done with a relatively small number of statisticians

Most survey are inventory and production type

Not a lot of demographic and financial expertise

FEDCASIC 2009

1997 NASS acquired the Census of
Agriculture

“The census is just a big survey”

or

“What do you do the other three years?”



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



All the Farmers in the US - 3,000,000+ mailout
Takes 22 months from beginning to end
64 FTE's from Census

Quotes from upper management about the Census

FEDCASIC 2009

Indicates a mind set

NASS Did Not Understand:

Scope and size

Complexity

Planning Required

Testing



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



Management very naïve about the process and size

Had never conducted a survey with a mail list of 3,000,000 records before

Nor published a report with this much data

Have not had to plan for or understand the issues with this size survey

Did not address testing strategies or time requirements

FEDCASIC 2009

Recap experiences of 1997, 2002 and 2007 Census of Agriculture



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



Management very naïve about the process and size

Had never conducted a survey with a mail list of 3,000,000 records before

Nor published a report with this much data

Have not had to plan for or understand the issues with this size survey

Did not address testing strategies or time requirements

FEDCASIC 2009

1997 Census of Agriculture

NASS acquired the Census of Agriculture shortly before mail out

Used existing systems from the Bureau of Census

Reengineered CATI in Blaise



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



Officially acquired in Feb 1997. Mail out 12 1997

Developed for the 1987 Census, improved in 1992, used the same system in 1997

The CATI process did not work in 1992. So NASS wrote their own Blaise instruments.

These worked well and have been used since with slight modifications

FEDCASIC 2009

2002 Redesigned the entire system

Mistakes:

Started Late

Took too much time to reorganize

Took on too much

Needed new system for editing and analysis

Tried too many new things

Should have focused on “need to have”



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



Re-organized and started working on the new process in early 2000.

Did not have enough knowledge of the process to know what would take the most time

Tried to develop every process from scratch.

Introduced scanning, nearest neighbor imputation, disclosure, database processing, UNIX processing,

Did not have a way of prioritizing the system requirements

FEDCASIC 2009

Underestimated staff resources needed

Census treated as collateral duty in subject matter areas

Edit development stalled

Removed collateral duties and relocated staff to stimulate progress

5 Branch Chiefs to Census Czar



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



Staff were asked to do their regular job in addition to the Census work
Operational activities always take priority to new stuff that is years away.
Project plans and milestones were created
Almost all milestones were missed
Teams became frustrated
Outside consultants were called in
One person was chosen to be the czar.

FEDCASIC 2009

Underestimated technical resources needed

Not enough testing time

Procured equipment too late

Not enough technical knowledge to
react to problems

Many layers of IT had to coordinate activities



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



Did not have much if any testing time

New UNIX machine arrived in November 2002. Became operational Feb 14, 2003.

No time to configure optimally before it was needed for production.

Learned on the fly.

Had the wrong DASD configuration

Did not have the proper tools and knowledge to monitor the box.

Difficulty working across infrastructure teams to identify possible problems.

No stress testing was performed at all.

Had to coordinate with UNIX, Communications, Database, Security, LAN, Developers, Business Users

FEDCASIC 2009

2002 Results:

Scanned OCR data had errors and false positives

Programs not fully tested prior to production

Database design flawed

Sybase & Redbrick combined

Did not play to their strengths

Unstable!



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



Respondents would draw line through the page indicating no response. However, the scanners would interpret as a 1, 7, 11 based on where the line hit the hot zones

Reran large amounts of data because of data integrity problems. Would move from one process to another and find out flags or values were not set properly. Would need to reprocess.

Had the 2 databases in the same partition. Did not play well together.

We would get POSIX locks. Forced us to reboot the databases once a week. The increased locks caused the databases to get slower and slower.

Didn't know that we needed to reorganize the databases on a regular basis when large amounts of data are being loaded.

Did not utilize the server side processing. Ran most things from the client.

Didn't allocate UNIX machine resources effectively.

Databases were unavailable about 40% of the time.

Batch processes would run into the daytime and slow down the day time processes.

FEDCASIC 2009

2002 Results: - Continued

Decision Logic Tables (DLT) Authoring Tool
operational

Delivered late, but was well tested upon
delivery

DLT capture from spreadsheets functional,
somewhat tedious

DLT compiler eliminated DLT programming



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



DLT tool was 6 months late. This short changed the DLT authors from being able to effectively test their logic.

Transferring the DLT rules from the spreadsheets was very time consuming. However once the data were in the database, these programs executed correctly.

DLT performance was not optimal

Data Review would sometimes have the "white screen of death"

Data was lost. Database tables did not stay in synch

FEDCASIC 2009

2002 Results: - Continued

No interactive edit (deferred batch)
DLT executables too slow
Nearest neighbor imputation developed
Quality varied by module
Publications delayed from 2/2004 to 6/2004
Still earlier than previous censuses

No process completely met expectations

Management tempered negative comments about
system and process



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



All records were edited in batch mode. Very little control on prioritizing edited or non edited records. Corrected records could take up to three days to be re-edited. It could take several editing sessions to get a record clean. Very frustrating for the end users.

Used a new nearest neighbor imputation strategy. This worked okay in the production and inventory sections of the questionnaire.

Didn't perform very well in the economic and demographic areas.

Every process experienced problems of one kind or another

Delayed the release of the Volume 1 publication from February to June. Still better than the Bureau of Census had done in the past.

FEDCASIC 2009

Management Changes Implemented from 2002 to 2007

Council of 10 mid level managers formed as 2002 results published along with 1 project manager

- Review planned changes to census process
- Monitor progress
- Recommend when resources needed adjusting
- Communicate plans to the rest of the agency
- Collected evaluations from users across agency
 - Was email application, now in central database



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



The Administrator felt the process needed more attention so 10 Branch Chiefs were put in charge.

NASS had a much better, clearer idea of what it would take to run the program now.

Basically, they made sure resources were available and had teams working as early as possible.

FEDCASIC 2009

Management Changes Implemented from 2002 to 2007 - Continued

Formed teams to recommend changes

Data capture through “clean data”

Nothing was off limits

Every process replaced, overhauled or refined

Gaps closed with new applications

Subject matter experts assigned full time to census & follow-on surveys

Used experience and lessons learned



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



The NASS staff that survived the 2002 program stayed with it for 2007. The experience provided much better understanding of what needed to be enhanced.

Time lines and milestones were much more realistic and able to be met.

FEDCASIC 2009

Processing Changes from 2007 (a few of many) and associated results

- Keyed from scanned images replaced OCR
 - Captured data as reported by respondent
- Key system components tested extensively
 - Few surprises once in production
 - Users touched system and learned applications before production phase
- New database design
 - Sybase and Redbrick played to their strengths
 - Sybase – transactional including edit
 - Redbrick – analytical
 - Replication process moved data from Sybase to Redbrick
 - Stable system (almost no unexpected downtime)



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



NPC has a system that only presented data to be keyed in marked cells and the area around that cell.

So there was no paper handling and the keyer is able to interpret the whole area while keying.

Had national training sessions with predefined test cases. Database was resettable.

Put each database in its own partition on the UNIX box.

Limited user to doing adhoc queries to the Redbrick database.

Interactive work and batch work ran concurrently.

Only had 3 unexpected down times in 12 months.

FEDCASIC 2009

Processing Changes from 2007 (a few of many) and associated results – Continued

DLT executable

About 75 times faster

Wide Area Network upgraded

Edit speeds “met” user needs

Interactive edit allowed users to learn how to work with edit

Imputation

Stratified donor pool on type of farm and value of sales

Edit made subroutine calls to imputation whenever need

Data Quality improved



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



In 2002, we learned that we didn't always get the best candidate record. The selection process needed to be refined so that we would find the distance of similar operations of similar size. We created a 2 level selection process. To accomplish this.

FEDCASIC 2009

Processing Changes from 2007 (a few of many) and associated results – Continued

Upgraded UNIX box and storage devices

Tested, tested, and tested some more

Every process improved

Service Level Agreement

Set expected number of records to handle daily

Database response and reliability

Benchmark processes

Quality control measures

Evaluate edit/imputation and data relationships

EDR

Sizing and integration

Macro aggregation – twice daily

Speed and timing



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



FEDCASIC 2009

2007 Results:

Published report on time in February 2009

For 2012 only need tuning

Ripple Impact of Census Effort

- Other systems delayed or postponed

- Census was first enterprise level editing/analysis system

- Survey application run on FO LANs

- Others on decades old mainframe systems



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



FEDCASIC 2009

2007 Results – Continued

Push to move other distributed applications to enterprise level

Survey Management (FoxPro)

CATI (Blaise)

Departmental initiative to secure servers in central locations

Support for distributed application very difficult



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



FEDCASIC 2009

2007 Results: – Continued

Team formed to decide data flow at new enterprise level

All survey processes involved

Frame maintenance

Sample selection

Instrument designers

Survey management (FO)

Data collection

Editing

Survey administration (HQ)

Analysis and summary

Output will be new business rules

Will be fed into database/application design process



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



FEDCASIC 2009

Future Direction of “Census” PRISM Processing System

Design considering survey needs
Tuning applications to meet survey needs
Team of end users to recommend specs
First census follow-on survey being developed



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



FEDCASIC 2009

Lessons Learned and Best Practices

Don't underestimate amount of communication required
Within teams & across various units

Manage requirements and scope creep

Leverage your experience & lessons learned

Have well defined milestones

Systems take longer to develop than planned



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)



FEDCASIC 2009

Lessons Learned and Best Practices

Incorporate quality control tools into the system

You can't do enough testing

Establish service level agreements

Define and manage expectations



U.S. Department of Agriculture
National Agricultural Statistics Service (NASS)

