Moving to a Centralized Database for Surveys in Blaise at the National Agricultural Statistics Service

Roger Schou, National Agricultural Statistics Service, USA

1. Introduction

The National Agricultural Statistics Service (NASS) is an agency of the United States Department of Agriculture. NASS is responsible for collecting, editing, and summarizing agriculture data. We are the sole agency for producing Agriculture Statistics for the United States. Our headquarters (HQ) is in Washington, D.C. and we also have forty-six field offices, six of which house CATI Data Collection Centers (DCCs). The capacity of the six DCCs range from to twenty to sixty workstations. Most of the other field offices also utilize a small staff of interviewers who collect data for their respective offices.

Over the past sixteen years, NASS has conducted surveys in Blaise in forty-four of our field offices. We have always collected and interactively edited data in a distributed environment. The collected data has been stored in Blaise datasets on each field office's LAN. In the late 1990's, NASS opened its first DCC which collected data for several client states. This introduced a complexity of moving zipped up datasets from one location to another. We have now grown to the six DCCs mentioned earlier. Since the data is disseminated, getting a quick snapshot of the progress of a particular survey on a national level is virtually impossible. We also rely heavily on the communication links being up considering how much data is physically transferred from a DCC to a client state.

During the last year, NASS has been thinking along the lines of centralized databases in our enterprise architecture. With the evolution of Blaise 4.81, we are able to pursue storing our Blaise data from all of our field offices in one central database. This will enable us to run national level reports quickly, as well as leverage this architecture to perform our pre and post survey activities. It will also eliminate the need to physically transfer files from one location to another.

2. A Centralized Environment

As we investigate moving Blaise data collection into a centralized environment, we have come to realize that this is not a quick, simple migration. Independently, it would be less involved; however, NASS has decided to move a number of its survey processes to a centralized environment, all within as close a timeframe to each other as possible. This includes centralizing sample master files, the NASS Survey Management System, our in-house web data collection tool (Electronic Data Reporting – EDR), Blaise data collection, our in-house Electronic Data Collection (EDC) tool for smaller surveys, and data analysis tools. Eventually our summary systems will be designed to read these database tables, as well.

What started out as a little splash of an idea to centralize one or two applications at NASS, has turned into a tidal wave on an enterprise level. This change has been welcomed by our agency and senior management has made it a priority. Open communication between all parties involved has been a key. A team has been created to examine Survey Processing in a Centralized Environment. It is made up of representatives from each of the sections involved to brainstorm and plan for this transition. We are trying to coordinate the transitions to the centralized solution so that as many pieces of the puzzle come together at the same time as possible. Many issues have surfaced as these meetings have progressed.

3. Database Decision

One issue that has been a concern is the database in which to store the data. NASS currently makes use of a number of databases including Sybase, Redbrick, FoxPro, Oracle, and MySQL. A proposal has been made to consider MS SQL as an alternative database to some of the existing applications and most of the new development. When centralized, we need as many systems as possible utilizing the same databases. The proposal is currently being reviewed and a decision has yet to be made. In the meantime, we have been instructed to use MySQL as the database for Blaise development, and our CASIC section has based preliminary development of the new system on this directive. A WEB-enabled menu, written in Visual Basic .NET, will be used to run the Blaise portion of the data collection and interactive editing. Blaise 4.81 will be our main tool for collecting and editing the data.

4. Generic BOI Files

We have decided to use Generic BOI files so that there is a common structure across all of the Blaise instruments. By storing the Blaise-collected data in the same eight tables, we envision only one translation tool

to extract the data from the Blaise MySQL database to the Work In Progress database (WIP) which will house the data from all of the data sources (CATI, paper, EDR, and EDC).

5. Versioning

We also plan on activating the Blaise versioning option. One of the requirements proposed by NASS senior management is that the original reported data must be preserved. This has not been the case at NASS in the past. By using versioning in Blaise, we will be able to identify the original data by date and time stamp. This will also allow us to do research on the amount of data that is actually being changed by editing.

6. Folder Structure

NASS has a future vision of opening two large DCCs to handle most if not all of its calling. Until that becomes a reality, we have to provide the ability to run CATI in the six DCCs as well as all of the field offices (FOs). Obviously, the CATI specifications will not be the same in all of the offices. Also, each office will need its own day batch. We have designed a folder structure to handle this challenge. Under each survey folder there will be an FO folder. Under this FO folder there will be individual folders for each field office conducting the survey. In each of these individual folders there will be a CATI specifications file (.BTR), a copy of the .BOI file, and a copy of the instrument files (.BDM, .BMI, and .BXI). We have yet to conduct testing to see if it is absolutely necessary to have a copy of the instrument files in each folder. It would be nice to only have to put them in one physical location.

7. User Access Rights

Another challenge introduced by centralization is user rights. In other words, what data can be seen and/or updated by each user. In the past, distributed instruments and data sets have by default enforced a certain level of user access security due to physical separation. The data has physically existed on different field office servers or in state-level SAS datasets on the Unix box, so the ability of individual field office staff to see data across states has been very difficult, if not impossible. When this data moves to a central database, user access rules will need to be enforced by a combination of database security and logic built into the applications. A preliminary policy has been written, and it has been reviewed and approved by management. A team has been assembled to discuss the details and to prepare for implementing this policy.

8. Additional CASIC Tables

8.1 Information Tables

The CASIC section has introduced three information tables in addition to the generic tables created by Blaise. A Survey Information Table will store needed information about an individual survey such as the folder name, the month and year, the instrument name, the .BOI file name, along with other characteristics of the survey. A Fips Allocation Table will store information such as which states (identified by numeric state FIPS codes) are responsible for the survey, which states are collecting the data, which states are editing the data, the start and end dates which correspond to the data collection period, along with some survey identifier fields. A User Information Table will store information about the users such as their names, employee numbers, and roles. These three tables are in their infancy stage. As plans develop for the other applications it is likely that much of this data will be shared, so the final table design will be determined at a later date. Early thoughts are to create a user access table on the fly from a combination of the information stored in these tables.

8.2 Error Limits Table

Error limits are a necessity for many NASS surveys. We carry limits for things such as acreage and yield for crops, maximum size for grain capacity, litter rates for hogs, maximum number of animals, and size of farm. Our CATI instruments flag warning and critical error messages to interviewers and editors based on these types of limits, which have historically been stored in external Blaise datasets. Error limits in the disseminated scenario have always been survey and state specific, and they have been stored in a Blaise data set unique to each survey in each state. Discussions are leading us to one centralized master error limits table that can be shared by multiple applications. Limits could be added for all states and all surveys. They could be stored, maintained, and accessed in this one location.

8.3 Previously Reported Data

NASS uses previously reported data (PRD) to control routing on subsequent surveys as well as determining an allowable percentage change from previous reports. PRD has been gathered and placed on the sample masters for each survey where it is needed. In a centralized environment, PRD should be accessible directly from the source where they are stored. Time will be freed up for the sampling people to work more with sample design and selecting samples by eliminating the busy work of extracting and posting PRD to the master files. It may be necessary to stage the PRD in a table, as there may be multiple sources for the data.

9. Initializing the Database

NASS maintains a list of farm operators from which survey samples are drawn. Once drawn, the names and addresses are extracted from our list frame. Four files are created from this process which store operator name and address, list frame comments associated with each operation, partners' names and addresses, if any, and sample master data. The current initialize process is a single click of a button that runs a large number of Manipula setups to check these input files, create folders, the external datasets, and ultimately the Blaise dataset.

In the centralized solution, the initialize process will be broken into three parts. The first part is the Survey Setup Process, which will be run by someone in HQ. It will create the folders on the application server and copy both the instrument files and the CATI specification file into each FO folder. The second part is the Initialize Preparation. This step will be run by the field offices. It does the checking of the name and address files against the sample master files. Any problems with the files should be dealt with by each field office, so they will be notified during this process of any issues. Once the files have passed these checks and balances, they will be placed in an area accessible by HQ and an indicator will be set that will signal they are ready for initialize. This brings us to the Initialize step. Most of the time, this will be a CRON job that will execute in the middle of the night to initialize the sample into the MySQL database. There will also be a button on the interface that certain HQ personnel may run for states that miss the deadline for the CRON job. When a state is initialized, another indicator will be set to indicate that they have been initialized.

People in our Survey Administration Branch will be able to keep up with the status of the states' progress during initialize as well as the whole survey process. Reports will be created to quickly show which states have completed the initialize process. This will allow HQ to quickly isolate any problems that may arise.

10. Testing Plans

The CASIC section has developed a testing plan to thoroughly test their system in the new centralized environment. We want to insure that we have not introduced anything that would cause the call scheduler to run less efficiently. In the United States, we must deal with data collection across several time zones. We don't anticipate any problems with the general concepts of the call scheduler, but we have changed the underlying principals to which we are accustomed and so we will plan to give it a thorough testing.

We will also load test the system. We plan on starting with ten interviewers in one location. As long as everything runs smoothly, we will then bring the number up to fifty interviewers in one location. The next step would be eighty interviewers in two locations. We plan on beginning with one instrument and then moving up to four instruments running at one time. The complexity of the instruments will increase as each one is added. On another night, we will have around 200 interviewers in three or four states using one complicated application. We plan to use both the scheduler and Manipula to retrieve cases. We may also want to start several sessions of the BtEmula.exe to get a larger quantity of traffic. Our final test will be during the day with up to 500 users including both interviewers and editors putting as much stress on the system as we might anticipate during our Census of Agriculture. We may also run several sessions of BtEmula.exe during this test phase.

The third test will involve user access. This testing will be done mostly from HQ. We will simply change our roles in the database to achieve different levels of access in order to test rights. Access needs to be restricted based on role and location of the user. For example, someone in the California office should not see data being collected and edited in the Florida office. Another example is an interviewer in one of the DCCs would need access to the state forms for which they are calling, but not forms that are the responsibility of another DCC.

11. State-Developed Blaise Applications

There are many instruments that have been developed by users in our field offices. A state-funded survey is an example of where the CASIC section in HQ would not be responsible for development or maintenance. We have always provided a hook into our HQ menu system for the state-developed Blaise applications. Once the national

surveys have been moved into the centralized solution, we will take a look at the state-developed surveys and explore the possibilities of once again providing a hook into the HQ system.

12. Benefits of Centralizing

We anticipate many benefits of moving to centralized databases for our survey processes. Real-time reporting of survey progress on a national level is something that we have always wished was available. Reports can be written that will be very helpful for the survey administrators. Eliminating the physical transfer of data files from one location to another is a huge step forward. Often, many hours are spent trying to locate data or recreate a transmission of data when a transfer was interrupted. Bringing all of our survey processes together so they may more efficiently communicate with each other will eliminate the “stove-pipe” applications that currently exist in NASS. When systems can be built with a service-oriented approach, the communications between the systems can be seamless.

13. Challenges

There are some challenges that we are proactively pursuing, while others will need to be addressed further down the line. Establishing and implementing a user access rights policy is one of those challenges. Another that is lurking is the coordination of a Blaise datamodel change. What are the steps needed to implement a new datamodel should the definition need to be changed after the start of data collection? What is the best way to communicate to users across the country that the datamodel has changed? How do we coordinate the downing of the CATI Service during data collection, if need be? Other challenges include coordinating multiple modes of data collection in central database. Many of our processes are moving to the centralized environment, but all will not get there at the same time. Dealing with the transition is a big challenge as bridges will need to be built to fill the gaps until all of the processes get centralized.

4. Conclusion

In conclusion, NASS is coming into some very exciting times of development. There is a very positive attitude towards centralizing the sampling, data collection, editing, analysis, and summary processes within the agency. This excitement is shared by the developers all the way up to the senior management. Although progress is slower than originally expected, NASS is making the correct decision by coordinating the transition of all of our survey processes. This will allow the processes to evolve with an inherited communication link between them. Many of the new features offered in the new design and technology can be leveraged when they can first be thought through. In the past, NASS would often place a high priority on turning out a product in a short amount of time and the full potential of a system was never realized. There will always be challenges when moving to a new architecture, but with open lines of communication and cooperation between developers, these challenges will be met.