Transforming Survey Processing at a Federal Statistical Agency

Nealon, Jack
National Agricultural Statistics Service, U.S. Department of Agriculture
1400 Independence Ave., S.W. Room 6054-South Washington, D.C. 20250
Jack.Nealon@usda.gov

Lehman, Joel
National Agricultural Statistics Service, U.S. Department of Agriculture
1400 Independence Ave., S.W. Room 6349B-South Washington, D.C. 20250
Joel.Lehman@usda.gov

I. Background: The success of our Agency's Long Range Plan is partially dependent on our successful transformation to Database-Optimized, Generalized, and Modular Survey Applications [1]. In the book, Good To Great [2], the author states: *"How a company reacts to technological change is a good indicator of its inner drive for greatness versus mediocrity. Great companies respond with thoughtfulness and creativity, driven by a compulsion to turn unrealized potential into results."* NASS has committed to this transformational effort to turn unrealized potential into cost savings and data quality improvements.

II. Objectives: There are three broad objectives to this transformational initiative. They are:

1. Reengineer or enhance applications to make optimal use of enterprise databases.
2. Retire redundant or duplicative applications so we operate more efficiently, e.g., fewer applications to maintain and fewer applications for employees to master.
3. Restrict standalone applications from being developed by providing generalized applications and enterprise databases for all surveys conducted throughout the Agency.

Once this transformation is complete, NASS will have a set of generalized, yet flexible, application services that will perform all processing tasks from survey inception to the publication of official agricultural statistics.

III. Benefits: Cost savings and data quality improvements will occur. For example, cost savings through staff reductions will be realized by implementing standard metadata, enterprise databases, and generalized applications since inefficient survey processing practices will be eliminated. Improvements in data quality have already been demonstrated when the Published Agricultural Statistics Database (called Quick Stats) was migrated to an integrated, enterprise database with standard metadata. NASS has already discovered and resolved about 5,000 data errors that existed historically in Quick Stats.

A more flexible, more integrated, more standardized, and more streamlined survey processing environment will ensure that cost savings and data quality improvements result. ***More flexibility*** will result since employees will be able to access applications and data anywhere in the Agency through thin-client or web-based applications working off centralized databases. This will allow our Agency to operate as one rather than 48 decentralized units since work can be shifted across offices to make optimal use of available staff resources. ***More integration*** will occur since standard metadata and enterprise databases will be shared across applications rather than applications having different data sources and metadata, as they often do now. These stovepipe applications have introduced work inefficiencies and data errors. In the future, integration will occur through standard metadata and centralized databases and not through multi-purpose and often complex applications interacting with proprietary and distributed data structures. ***More standardization*** will be provided by having standard survey procedures and generalized application services for all surveys, which will reduce staffing requirements and provide opportunities for data quality improvements. Finally, ***more streamlining*** of survey processing is being pursued at NASS. In the 2010 edition of SMART Enterprise [3], the author states that: *"MRI's have shown that the brain actually fires differently when people are doing the simple process of adding numbers, as compared to subtracting them. It's because we are born to add, collect, hoard, and consume. The trick, then, is to understand what to eliminate."* The following five examples are streamlining efforts already underway:

1. There will be no need to create, manipulate, and transfer thousands of data files, such as Blaise and SAS files, from application to application when the data resides in a centralized database.
2. The number of processing platforms will be reduced. The enterprise databases that support our survey and census data processing will be hosted on the UNIX/Linux platform along with some enterprise software, such as widely-used SAS. This will eliminate the need for NASS to also continue to process survey data on Mainframe and Windows platforms.
3. Fewer generalized national applications will be needed. For example, NASS has had two call schedulers and two CATI systems when we should only invest in one call scheduler and CATI system.
4. The number of customized applications in our Field Offices will be minimized by providing a set of generalized national application services for Field Offices to use.
5. Manual work activities will be reduced through more automation, such as implementing significance editing and automating more list frame maintenance tasks.

IV. Technical and Business Solution: *Standard metadata* will be used across our entire agricultural statistics program. *Transactional databases* (optimized for capturing and updating records) *and analytical databases* (optimized for accessing records for analysis and generating reports from the analysis) will use the standard metadata. *Application services* or modules, such as an imputation service, will interact directly with the enterprise databases. A brief description of each of these three critical components will now be provided.

A. Metadata: Metadata will form the hub for all data processing. Metadata provides the "who, what, when, and where" for every data item that we collect and store. Currently, metadata can change from application to application and from survey to survey, which not only has introduced work inefficiencies, but also has made the process more susceptible to data errors. Therefore, metadata standards are being developed and enforced across all surveys. The three critical levels of our metadata are: (1) enterprise metadata, such as the list of survey names and all master survey variable names, (2) survey-specific metadata, such as survey time and survey key or item codes, and (3) state-specific metadata, such as specific state conversion rates or rounding rules.

B. Enterprise Databases: The major enterprise databases have already been designed and some are already operating in a production environment. There are two types of database designs utilized at NASS: On-Line Transaction Processing (OLTP) and On-Line Analytical Processing (OLAP). OLTP databases are characterized by a large number of short on-line transactions (retrieve, insert, update, and delete). The main emphasis is on transaction speed and data consistency. An example of an OLTP database at NASS is our sampling frames database called ELMO (Enhanced List Maintenance Operations). On the other hand, OLAP databases are often characterized by retrieving a large number of records for aggregation, analysis, and developing reports based on the analysis. The main emphasis is on retrieval speed of multiple records. Our 8-billion record Data Warehouse is an example of an OLAP database. A brief description of each of the enterprise databases follows. Figure 1 provides a graphical depiction of the NASS enterprise databases. The databases shown in *italics* are the OLAP databases.

(1) Pre-Survey Activities: The ***Metadata Repository database*** has already been developed and is being used to provide standard metadata descriptions to the Question Repository database and aggregate metadata to our aggregate applications, such as the Publication Tool (called PubTools). The ***Question Repository database*** will contain all survey and census questions used in NASS surveys. It serves as the source for the design and building of survey questionnaire instruments for paper and pencil interviewing (PAPI), computer-assisted web or self interviewing (CAWI or CASI), computer-assisted telephone interviewing (CATI), and computer-assisted personal interviewing (CAPI). When the Metadata Repository database is updated, the relevant updates are automatically posted to the Question Repository database.

The ***Sampling Frames database*** contains individual farm, ranch, and agribusiness information used for sampling and survey purposes. This database also catalogues the tract operators in the area frame sample, as well as serving as the repository for the list of subscribers to NASS publications. This database is referred to as ***ELMO*** (Enhanced List Maintenance Operations). Changes are being made to ELMO to support the list frame activities in the National Operations Center (NOC) to be opened by NASS in St. Louis, Missouri during August 2011. The NOC will provide some centralized survey services, such as list frame maintenance and CATI.

The ***Survey Management database*** is a new database, and will focus on survey preparation, management, and coordination. For example, information on each individual in the sample for a survey, such as personal identifiers and sampling weights, will be part of the Survey Management database. Information needed to effectively prepare for survey interviews will be provided, such as the data collection strategy (mode of data collection for individuals) for creating the survey instrument.

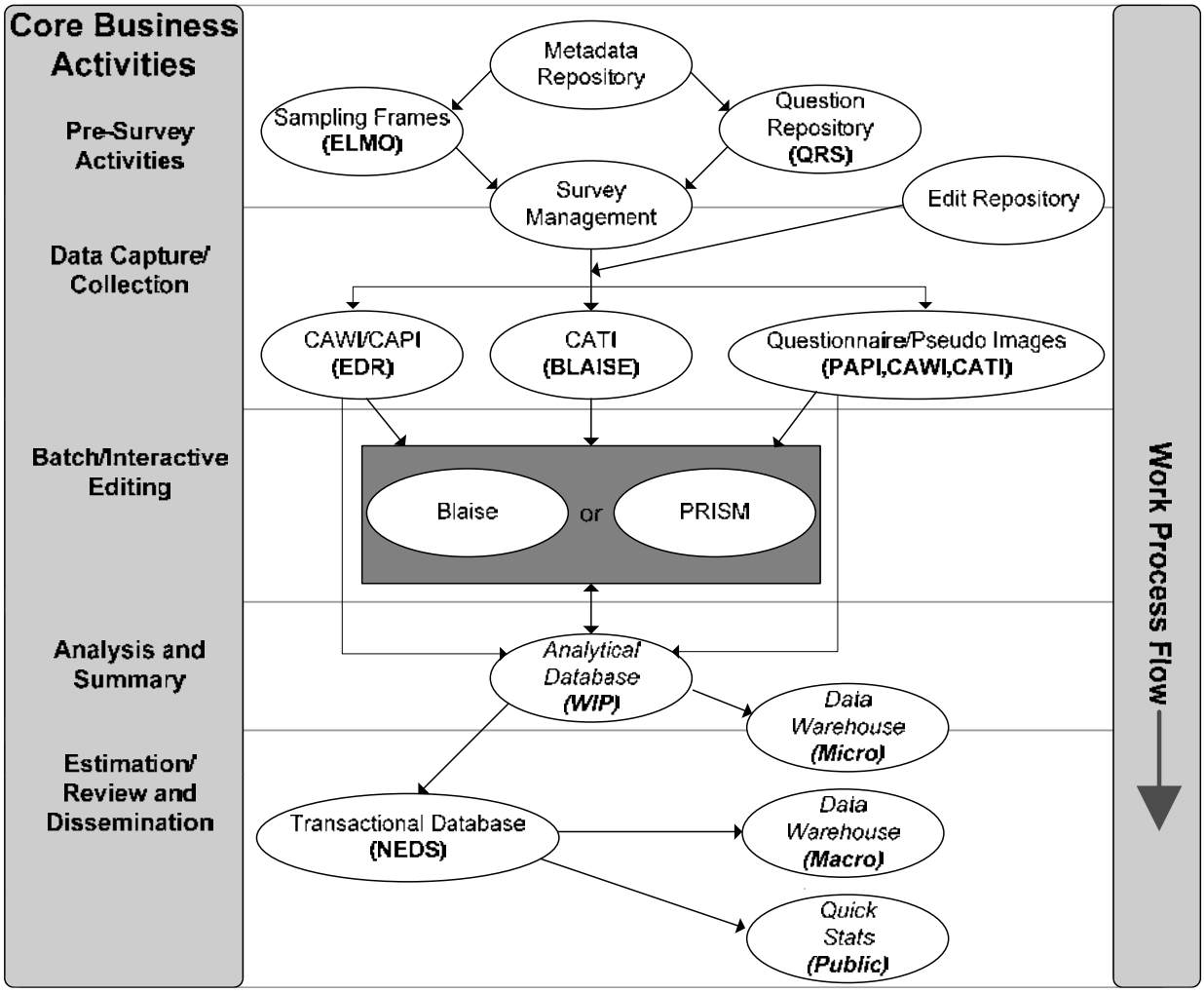
The ***Edit Repository database*** will store edit logic to ensure data consistency and edit limits (upper and/or lower limits) to ensure data reasonableness. The goal is for the Edit Repository database to feed the edit logic and edit limits that are needed for CAWI, CATI, CAPI, and any edit or imputation application.

(2) Data Capture/Collection: Data collected and edited using the CATI and editing system called Blaise from Statistics Netherlands will change from a distributed data environment using proprietary Blaise data sets for each survey to a **centralized database for Blaise data**. This transformation will provide significant efficiencies to our survey processing since Field Offices will no longer need to create Blaise input files and Blaise data files for every survey and distributed LAN updates will no longer be necessary. The plan is to use this data capture centralized database to capture all data collected through Blaise in the future, regardless of whether the data originates from CATI, CAPI, or CAWI.

Data collected on the Internet using the ***EDR*** (Electronic Data Reporting) application at NASS is already captured in a ***centralized database***. The EDR centralized database is already used for CAWI in 252 surveys and censuses and for CAPI testing in several Field Offices.

The ***Questionnaire/Pseudo Images database*** will contain snapshots of each page of a survey or census questionnaire submitted by paper and pencil interviewing (PAPI). It also will contain generated images or pseudo-images of CATI, CAPI, and CAWI responses.

Figure 1: NASS Enterprise Databases



(3) Batch/Interactive Editing: The data captured through CAWI, CATI, CAPI, and PAPI will be stored in the ***PRISM or Blaise CATI/Editing transactional database*** for the data editing process. PRISM is an acronym from a decade ago meaning Project to Reengineer and Integrate Survey Methods. The PRISM database, like the Blaise CATI/Editing database, contains individual farm, ranch, or agri-business data for the purpose of edit processing. The PRISM and Blaise databases are not the same because the PRISM data structure is more complicated and different from the Blaise data structure. These databases store the final edited data.

(4) Analysis and Summary: The ***analytical database, called WIP (Work In Progress),*** will be a single, high performance OLAP database for all census and survey data and will contain multiple years of data in addition to survey data currently being processed. This database will provide the current and historical data for editing, analysis, and summary/tabulation processes. The WIP analytical database will also contain the information needed to manage surveys (such as check-in information), track the disposition of collected data, and generate a variety of management information reports. After the survey is completed and the official agricultural estimates published, the micro-level data will be loaded to the ***Data Warehouse***. This easily accessible database contains the deep history of farm, ranch, and agri-business data and currently has available survey and census data from 1997 through 2011.

(5) Estimation/Review and Dissemination: The ***transactional database called NEDS*** (National Estimates Database System) contains the information required to either interpret multiple survey indications and other inputs, such as administrative data, to set or derive an official estimate, or to review survey indications for reasonableness and approval prior to dissemination. This transactional database contains the current aggregate information as well as historical aggregates needed for analysis and review, e.g., past ten years. This database has already been designed. The historical

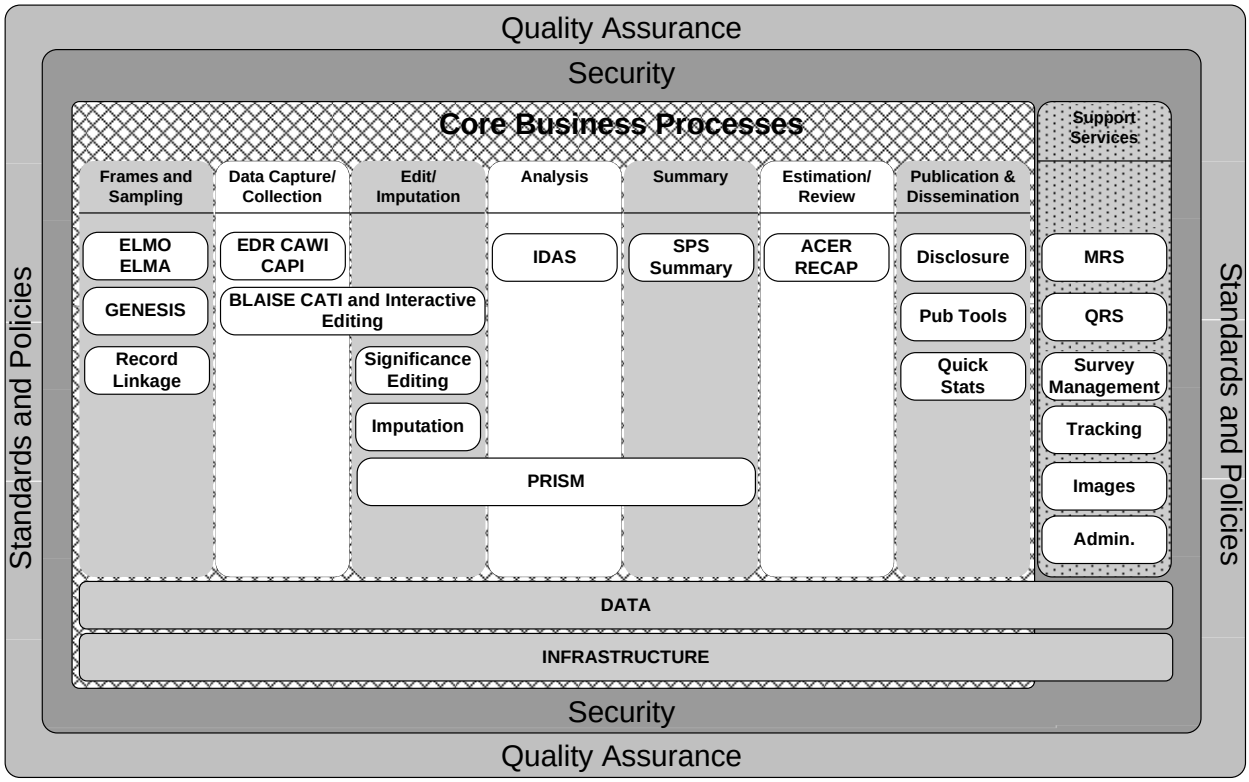
aggregate information will also be loaded to a ***Data Warehouse*** which will be the analytical database containing all historical survey indications, check data, and estimates for additional analysis and research purposes. This database, called the ***NEDS Analytical database***, is already partially populated with historical information.

The ***Quick Stats database*** was deployed for public use in 2010 and integrates all published official agricultural estimates into a simplified database schema containing 5 tables utilizing hierarchical metadata for simplified browsing. The Quick Stats database will serve as a single data source for all of NASS' published estimates.

C. Application Services: In an online survey in 2010 [4], 166 Federal information technology leaders generally agreed that "*application modernization is a "must-do" excursion to sustain services in an era of strained budgets."* Figure 2 summarizes the application services at NASS that have been or are being modernized to work optimally off the enterprise databases. We will now briefly describe these application services for NASS's support services and seven core business processes.

(1) Support Services: The ***Metadata Repository System*** **(MRS)** is a thin-client, web based application providing a set of role-based browsing, loading, and editing tools for metadata. These tools ensure that governance rules are enforced and greatly aid stewardship. The ***Question Repository System (QRS)*** is an application providing a structured environment to create survey questions and combine those questions into complete survey instruments. Survey instruments created in QRS can be directly utilized for PAPI, CAWI, and CAPI data collection. Current development is underway to generate Blaise's CATI instruments directly from QRS with minimal human intervention. The ***Survey Management*** system will be very different in a centralized database environment. The current multi-purpose Survey Management System (SMS), which is a currently distributed file based system, will be replaced with a series of application services accessing centralized databases.

Figure 2: NASS Application Services



There will also be additional support services developed primarily to support the NOC in St. Louis, Missouri. The ***Tracking and Control*** application will replace the existing check-in features found in the Survey Management System. The application will be used to check in mail receipts, move work among processing units, and monitor the workflow. It will also provide an audit trail at the individual record level and generate management information reports using the WIP analytical database. The ***Key From Image*** is an application designed to automate the data entry process by only presenting fields with entries in data fields. The Key From Image service will be flexible and allow for a mixture of heads-down data entry and key from image. The **Pseudo Image Creation** is an application that will present captured data electronically, e.g., CATI and CAWI, in a paper questionnaire format that is viewable on screen.

(2) Frames and Sampling: The ***Enhanced List Maintenance Operations* (*ELMO*)** consists of multiple application services for retrieving information from the ELMO database. Examples of these application services are simple fact checking on an operation, updating a telephone number, and extracting name and address information. These applications already operate effectively within a centralized database in the UNIX environment. The ***Enhanced List Maintenance Assistant* (*ELMA*)** is an application recently developed to support the centralized list frame activities from the NOC. ***GENESIS (Generalized Enhanced Sampling Information System)*** is the Agency's sampling application. This SAS application is designed primarily to define sampling populations and select samples for surveys. GENESIS works with the Data Warehouse and ELMO to access data from multiple previous surveys for sampling purposes, assign sample sizes and sample weights, and select the samples storing results in SAS datasets. A redesign of GENESIS will focus on migrating from SAS datasets to a centralized database. Finally, the ***Record Linkage*** application is used to remove duplication from list sources by identifying records believed to correspond to the same entity. This application already operates successfully off a centralized database.

(3) Data Capture/Collection: The ***Electronic Data Reporting (EDR)*** application utilizes survey instruments created directly from the Question Repository System for web-based data collection (CAWI). Since the EDR application is data driven from centralized databases requiring minimal human intervention, NASS is currently able to conduct 250 different web based surveys annually. NASS has recently been able to utilize the thin-client, database optimized flexibility of EDR as the basis to provide data collection instruments for CAPI on Apple Ipads. This method of CAPI data collection is currently being tested in several Field Offices. At the time of the initial testing, Blaise CAPI did not satisfy these requirements so is not part of the initial CAPI testing. However, NASS plans to explore the potential use of Blaise as its future CAPI application.

NASS has been using ***Blaise CATI*** software since the beginning of the 1990's. Blaise is a file based, thick client application and is distributed across our Field Offices on the Local Area Networks (LANs). NASS has recently been centralizing and consolidating the file and print servers from 48 locations across our Agency [5] and will deploy Blaise CATI in a centralized database environment. This will be much more efficient since there will no longer be the need to create, manipulate, and transfer thousands of Blaise files and remove the need to distribute software/instrument updates to the Field Offices when a change is made to a Blaise application. We also plan to create an interface from the Question Repository and Edit Repository to Blaise so that Blaise CATI/editing instruments will be efficiently generated and maintained in the future for all NASS surveys using telephone data collection. These efficiencies will also allow NASS to retire an internally-developed CATI/editing system called Electronic Data Collection (EDC). Blaise has the potential to provide NASS a single, integrated CATI, CAPI, and CAWI system sometime in the future. Blaise CAWI is not well integrated with Blaise CATI/CAPI now, but is expected to be when the next generation of Blaise (Blaise 5) is completed around the end of 2012.

(4) Edit/Imputation: The Edit Repository database will store generalized edit logic and edit limits, which will eventually be used by ***EDR/CAWI, CAPI, Significance Editing, Imputation, Blaise CATI/Editing, and PRISM editing.*** The PRISM edit system, which originally serviced the Census of Agriculture and Census Follow-On Surveys, is being enhanced with multi-layer editing capability to accommodate field-level editing on the June Area Survey and application-level editing on the Chemical Use Surveys. Surveys currently using the NASS Survey Processing System (SPS) Edit will migrate to PRISM or Blaise editing so that the SPS edit will be retired. The generalization, centralization and reduction in editing systems (SPS,EDC) mentioned above will provide NASS with additional resource and infrastructure efficiencies.

To further increase operational efficiencies, NASS is researching the potential for streamlining the labor-intensive, manual edit/review process for many of our Agency's surveys. ***Significance editing and automated outlier detection*** distinguish responses likely to have a large impact on the survey results from those responses that could appropriately be automatically corrected (as needed) through statistical editing/imputation applications. This process enables the manual review and correction process to focus on ensuring the quality of "impact" reports rather than spending time reviewing all edit-failing reports. The core software for this significance editing and automated outlier detection approach is Banff from Statistics Canada. A parameter-driven ***Imputation*** application is also being developed using the Banff system.

(5) Analysis: The ***Interactive Data Analysis System*** (IDAS) is a distributed, LAN-based, thick-client SAS system. Each instrument is custom designed for a survey. IDAS is a very valuable analysis tool, but is not an operationally efficient application. It will be much more efficient when we develop a generalized set of IDAS analysis views (called ***GIDAS***), specifically for the 153 surveys to be standardized in our Agency, and retool IDAS to work optimally off the centralized WIP analytical database rather than countless distributed SAS files. The implementation of GIDAS will start in May 2011 using the WIP centralized analytical database.

(6) Summary: The plan is to expand the usage of the twenty-year old ***SPS Summary*** (written in SAS) from the larger, national surveys to the smaller surveys. Our multiple estimators are already coded, debugged, and serviceable in the SPS Summary. The SPS Summary will also work directly with the WIP analytical database. We have developed an interface to make it easy and efficient to create summaries for many small surveys. The SPS Summary will then feed the survey indications directly to the NEDS transactional database.

(7) Estimation/Review: For a number of NASS surveys, such as crop and livestock surveys, employees analyze aggregate information based on surveys and administrative programs using multiple survey indications, administrative data, historical time trends, balance sheets, and commodity expertise to amalgamate the information into the official estimate. Currently, antiquated Formula1 and Lotus123 spreadsheets are used in a distributed file based environment for this estimation process. A lack of standardization and little audit control combined with the hundreds of files generated by this process create an environment that is inefficient and prone to data errors. The NASS Estimation application in the future (currently called ***ACER*** for Analysis, Comments, Estimation, and Review) will be a generalized application integrated with the NEDS centralized database.

For about 332 surveys, NASS publishes the survey indications that are generated from the summary system. Therefore, an ACER estimation application is not needed. Instead, a generalized set of review screens will be developed for the 332 surveys where employees will only review the survey indications and comments from the Field Offices and/or Headquarters, and then either approve the estimate for public dissemination or request additional micro-data analysis be conducted before approving the estimate. This application is called ***RECAP*** (Review Estimates & Comments, Approve & Publish). A single graphical user interface and single point of entry will be used for accessing commodities or items for review (RECAP) or estimation (ACER).

A SAS-based application is used to run ***Disclosure*** on the Census of Agriculture and selected other surveys and censuses. This application uses the p-percent rule and through a series of complex algorithms identifies primary suppressions that fail the rule. The application also identifies complementary suppressions to prevent primary suppressions from being derived. Employees also identify linear relationships from one published table to another, and these relationships are incorporated into the Disclosure application to prevent previous suppressions from being derived.

(8) Publication and Dissemination: The use of the ***PubTools*** application will be expanded to create official reports or releases not only for the larger, national releases, but also for smaller, Field Office releases. The existing PubTools application has already been modified to access both the NEDS and Quick Stats centralized databases. However, the ease of use is also being improved by having PubTools accessed through a simplified graphical user interface.

The ***Quick Stats*** application was implemented in 2010 and leverages the integrated Quick Stats database. The Quick Stats application provides: "Build Your Own" ad-hoc query capability, "Data By Subject or Commodity" query capability, "Keyword Search" query capability, and "Pre-Defined" (or Pre-Designed or Canned) queries. The Quick Stats application provides improved functionality, such as the ability to query across commodity sectors, e.g., corn and hogs, to pivot results (switch rows to columns), to show or hide columns in the results table, and to export the query results. The plan is to develop hundreds of pre-defined queries or tables on crops, livestock, demographics, economics, et cetera to meet the diverse needs of our data users.

V. Implementation Timeline: The following quote, in an edition of SMART Enterprise [6], summarizes our implementation strategy: *"Don't boil the ocean. Don't blast one huge project out there that promises to change everything…it is best to have lots of small, quick wins on new projects."* Therefore, NASS is taking an incremental approach to this initiative where most of the transformation will be completed during the next two years (by September 2013) with the remaining work completed within three years (by September 2014).

VI. Conclusion: This transformational initiative to Database-Optimized, Generalized, and Modular Applications will provide cost savings and data quality improvements. How significant these cost savings and data quality improvements are will depend on the level of commitment and support from employees throughout NASS for this initiative. NASS is a great organization so the expectation is that employees will continue to embrace this transformational initiative to finally turn unrealized potential into significant improvements for NASS products and services. With tighter federal budgets being forecast, the operational efficiencies realized from this transformational initiative will position NASS to continue to provide many valuable products and services in service to agriculture.

VII. References:

[1] Nealon, Jack, Transformational Initiative #2: Database-Optimized, Generalized, and Modular Applications (DOGMA) – Vision and Implementation Plan, National Agricultural Statistics Service, U.S. Department of Agriculture, December 2010.
[2] Collins, Jim, Good To Great, HarperCollins Publishers, Inc., 2001.
[3] Hulme, George V., SMART Enterprise, 2010.
[4] Bass, Brad and Tobin, Mary, Federal Application Modernization Road Trip: Express Lane or Detour Ahead, UNISYS and MeriTalk, January 2011.
[5] Gleaton, Elvera, Transformational Initiative #1: Centralizing LAN Services – Vision and Implementation Plan, National Agricultural Statistics Service, U.S. Department of Agriculture, May 2011.
[6] Petrisko, Michael, SMART Enterprise, 2010.