



Transformational Initiative #2

Database-Optimized, Generalized, & Modular Applications (DOGMA)

Vision and Implementation Plan

Jack Nealon
National Agricultural Statistics Service

December 30, 2010

Research, Education and Economics (REE)

U.S. Department of Agriculture (USDA)

Table of Contents

Background.....	3
Objectives.....	5
Benefits	5
Technical and Business Solution	7
Metadata	8
Enterprise Databases.....	9
Application Services	14
Critical Success Factors.....	22
Implementation Timeline	23
Conclusion	25

I. Background: The success of our Agency's Long Range Plan is partially dependent on the successful transition to Database-Optimized, Generalized, and Modular Applications. For example, the development of customized survey questionnaires will not be realized unless NASS implements a centralized database environment. The ability to disseminate customized and more diverse data products will not be realized unless NASS fully leverages the recently released Published Agricultural Statistics database on the NASS web page (called the Quick Stats database). NASS will not be positioned to significantly strengthen its analytical capability if we do not migrate to easily accessible centralized databases. Finally, the vision of our Administrator, Cynthia Clark, to improve data quality, produce cost efficiencies in the production of agricultural statistics, and position NASS to serve the mission of USDA more broadly will not be fulfilled without this transformational initiative being successfully implemented.

NASS has several challenges to overcome so our applications operate efficiently in an enterprise metadata and database environment. While operating in centralized databases provides a number of benefits in itself, there are other important benefits to be gained, such as streamlining the processing environment, upgrading our applications, and transitioning all surveys to more efficient and effective processing. The result will be improved data quality, data availability, and data security. Using common metadata will ensure consistency within and across survey programs. Migrating to modular applications (in lieu of multi-purpose and often complex applications) will reduce the risk of applications often becoming obsolete as they have in the past.

At the NASS National Conference in San Diego in April 1995, the Data System 2000 Steering Committee, chaired by Fred Vogel, set forth a vision for NASS to operate in a centralized database environment rather than our distributed environment using proprietary data structures, such as FoxPro, Blaise, and SAS. The NASS Strategic Plan 1994 stated: ***“The mission of NASS requires effective and efficient movements of data to NASS, within NASS, and out of NASS to customers. This strategic initiative (NASS Data System 2000) addresses how the Agency will manage its data in the future.”***

Our first successful efforts with centralized databases occurred during the second half of the 1990's with the Enhanced List Maintenance Operations (ELMO) for list frame activities and the Data Warehouse for accessing historical survey and census data. For example, Figure 1 illustrates the great success NASS has had over the years making available to employees in the Data Warehouse over 7 billion historical survey and census data responses. Usage statistics for the Data Warehouse show that over 98 percent of the queries against the database resolve in less than two seconds and the general nature of the queries has been to select or sum a small amount of data.

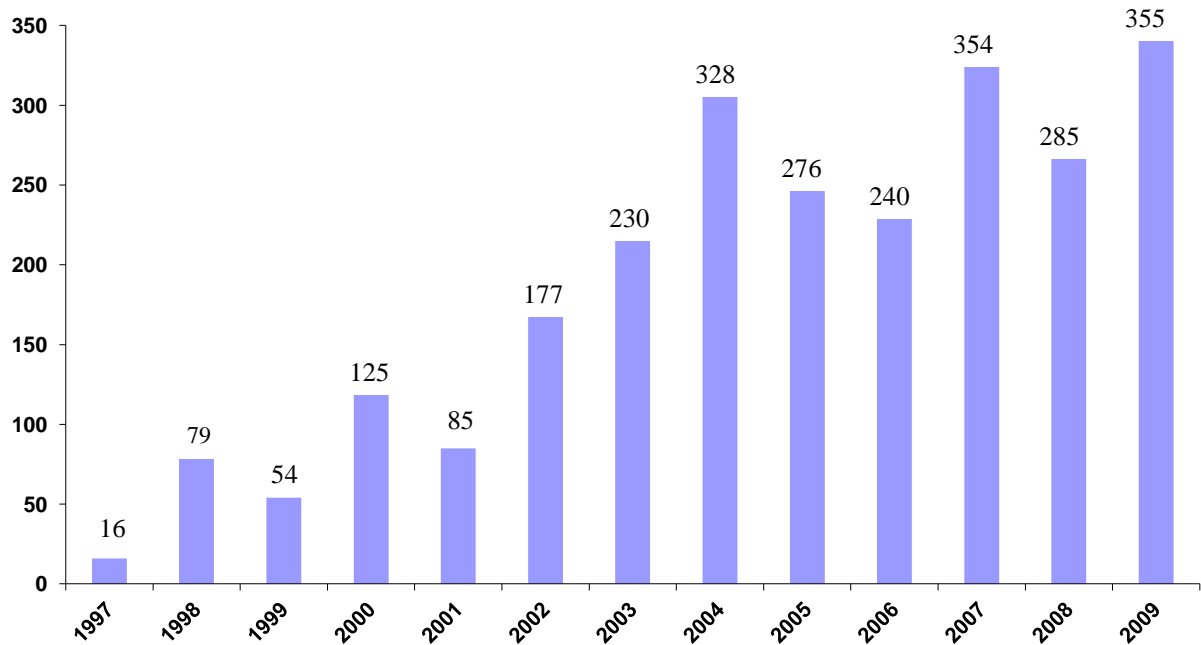
Despite these early successes, the vision of Data System 2000 stalled for over a decade due mainly to the following five reasons:

1. Significant developmental staff resources were focused from 1997 to 2007 on developing a high performing Census processing system, which included the optimal use of a 32-

processor UNIX server, transactional and analytical centralized databases, and Census SAS applications working effectively with relational databases.

2. Our annual survey program expanded in scope since 1997, which diverted even more staff to supporting our ongoing programs. For example, the number of reports published by NASS increased over 50 percent from 1997 to 2010 (333 reports on the Agricultural Statistics Board calendar in 1997 compared with 512 reports in 2010).
3. NASS did not have sufficient expertise in transactional and analytical databases until former NASS Administrator, Ron Bosecker, authorized the formation of the Data Services Branch in 2006, and then supported staffing the new branch with the necessary database expertise to advance multiple mission-critical enterprise databases.
4. The use of centralized databases and standard metadata were not widely embraced since application development and maintenance can be more convenient and independently serviced if the data and metadata are controlled by the application. But this approach lacks an enterprise view since the entire survey process is often not optimized.
5. Information technology is constantly changing so employees need to learn new skills and embrace technological changes that benefit the Agency. Skills have lagged, for example, in how to make optimal use of centralized databases and in developing thin-client applications.

Figure 1: Number of Surveys Loaded into the Data Warehouse Each Year (1997-2009)



In the book, *Good To Great*, by Jim Collins, he 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.”*** As NASS recommits to the vision of the Data System 2000 Steering Committee, we need to allocate the appropriate staff resources, utilize contractor expertise wisely, collaborate fully with one another, reengineer our LAN-based applications, and provide sufficient training dollars, particularly for training in the effective use of databases. We will then be able to deliver cost savings and improve data quality by providing a more flexible, more integrated, more streamlined, and more standardized processing environment. If NASS embraces this transformational initiative, then the cost savings and data quality improvements will be significant.

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

1. Reengineer or enhance applications to make optimal use of enterprise databases.
2. Standardize about 153 non-standard surveys across our Field Offices and use generalized applications and enterprise databases for these surveys.
3. Reduce the number of applications throughout the Agency so we operate more efficiently, e.g., fewer applications to maintain and fewer applications for employees to master.

Once this transformation is complete, NASS will have a set of application services that will perform all processing tasks from survey inception to the publication of official agricultural statistics. The numerous applications and processing approaches that have evolved across our decentralized Agency during the past two decades will be reduced to a single processing approach using standard metadata and enterprise databases with various application services available for a particular survey.

III. Benefits: Cost savings and data quality improvements will occur. For example, Carol House, the previous senior project manager for this initiative, estimated in February 2010 that the cost savings to be realized from implementing database-optimized, generalized, and modular applications could total about \$3.1 million per year or a savings of about 30 staff positions. Significant additional staff savings should also be realized from the expected widespread implementation of selective/statistical editing discussed later in this document.

Improvements in data quality were recently 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 had existed historically in Quick Stats. Therefore, a more flexible, more integrated, more streamlined, and more standardized 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.

Work can be readily shifted from one Field Office to another Field Office, from Headquarters to a Field Office, or from a Field Office to Headquarters, which will allow us to make optimal use of available staff resources. For example, work can be optimally distributed to improve response rates and resolve data anomalies in a timelier manner.

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. For example, there is a cross reference file to redefine or link metadata between the SPS Summary and NEDS FoxPro files. There is also a cross reference table that not only resolves inconsistencies in the metadata between NEDS and Quick Stats, but also defines some metadata that is not provided by the SPS Summary or NEDS, i.e., units of measure. We plan to stop allowing different applications and data stores to have different metadata. 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 streamlining will also provide significant benefits. In the 2010 edition of SMART Enterprise, George V. Hulme stated: *"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."* NASS certainly has opportunities to streamline our processing environment. The following five examples are streamlining efforts that will eventually occur:

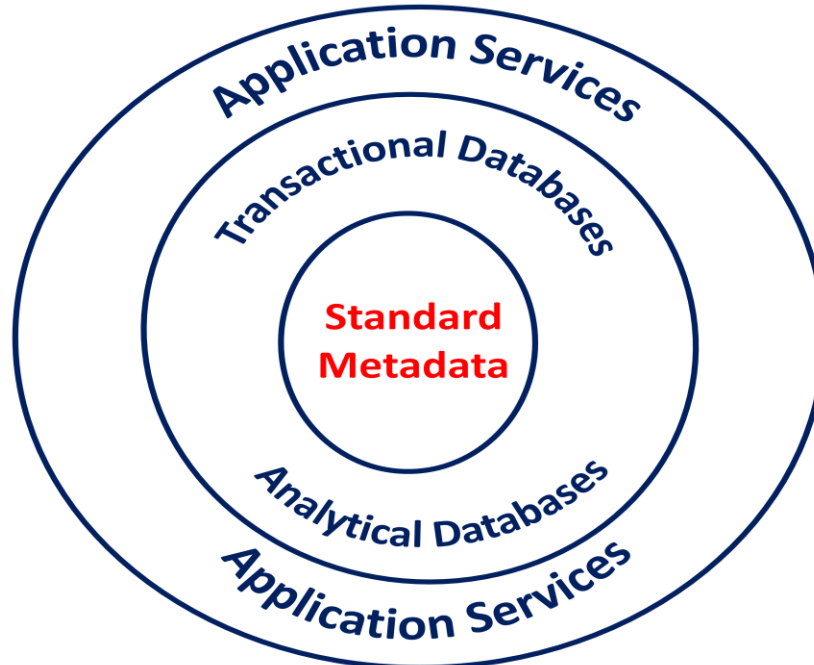
1. There will be no need to create, manipulate, and transfer thousands of data files, such as Blaise and SAS files, from application to application when the data resides in a centralized database. This will significantly streamline survey processing activities.
2. The number of processing platforms will be reduced from four (Mainframe server, UNIX/Linux servers, Windows servers, and Windows desktops) to two (UNIX/Linux servers and Windows servers), which will eventually simplify our processing. 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 to process on the mainframe when we fully migrate to the UNIX/Linux platform for our survey and census data processing. When thick-client applications are reengineered to thin-client applications, Windows desktop data processing will disappear.
3. Fewer generalized national applications will be needed. For example, NASS has two call schedulers and two CATI systems when we should only invest in one call scheduler and CATI system. We also have multiple applications doing the same types of edit functionality so we will phase out duplicative editing systems.
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 the selective editing service and automating more list frame maintenance tasks.

More standardization will provide cost savings and opportunities for improving data quality. Having standard metadata for the entire survey or census process minimizes the introduction of data errors. Having standard survey procedures and standard database-optimized application services across Field Offices for the 153 non-standard surveys will certainly reduce staffing requirements and provide opportunities for data quality improvements. Standardization also can reduce the complexity of survey processes and reduce training demands on employees.

IV. Technical and Business Solution: As depicted in Figure 2, **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 or a selective editing service, will interact directly with the enterprise databases. A description of each of these three critical components to our technical and business solution, namely, metadata, databases, and application services, will now be provided.

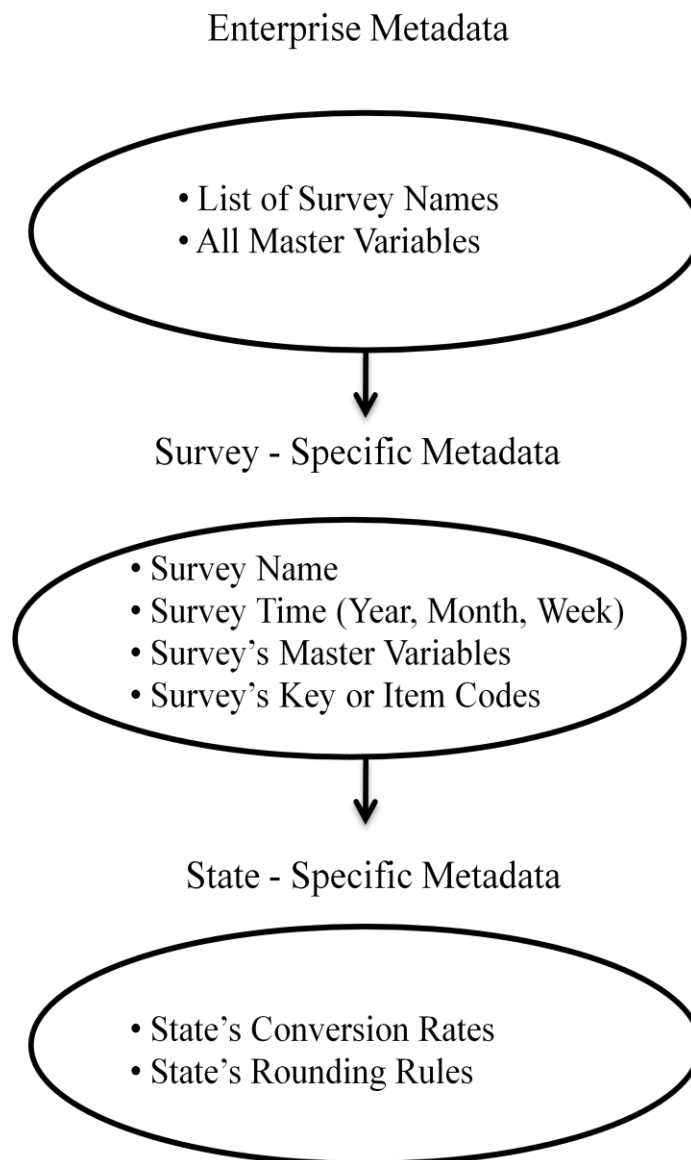
Figure 2: Key Survey Processing Components



A. Metadata: Metadata will form the hub for all data processing. Metadata is “data about data.” It provides the "who, what, when, and where" for every data item that we collect and store. Therefore, it helps steer our data-driven processes. Metadata is structured information that describes, explains, locates, or otherwise makes it easier to describe, identify, retrieve, use, and manage our data.

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 will be enforced across all surveys for items such as variable names (e.g., CCRNXXPL is corn planted acres). The three critical levels of our metadata are shown in Figure 3.

Figure 3: Critical Levels of Our Metadata



B. Enterprise Databases: The major enterprise databases have already been designed and some are already operating in a production environment. This transformational initiative will utilize the NASS Database Standards approved by the NASS Enterprise Architecture Council, the NASS Business Council, and the NASS Senior Executive Team.

Our enterprise databases that support our survey and census data processing will be hosted on the **UNIX/Linux platform**. The UNIX/Linux platform provides the performance, scalability, and security needed for our enterprise databases and our survey and census processing. NASS has invested significantly in UNIX/Linux services, has considerable experience and expertise in this area, and has multiple highly successful deployments in the UNIX/Linux environment so NASS will maximize its usage with this transformational initiative.

There are two types of database designs: 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. To achieve this objective, the database design usually consists of a highly normalized data model often with many tables, and focuses on optimizing single record processing time. 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. To achieve this objective, the database design is typically denormalized with few tables to optimize access time to multiple records. The Data Warehouse is an example of an OLAP database.

Sybase has been our OLTP database management system since the development of ELMO in the mid-1990's and is now owned by SAP. However, NASS has pursued a more economical solution during recent years, where applicable, using MySQL as our OLTP database, which has saved a considerable amount of money. MySQL, which is now owned by Oracle, is being used successfully by many organizations, such as Ticketmaster, Craigslist, Yahoo, Associated Press, Evite, NASA, Los Alamos Labs, Google, Wikipedia, and YouTube. Red Brick, which is now owned by IBM, is NASS's sole OLAP database management system. SAS, which NASS uses very broadly for sampling, editing, imputation, analysis, summary, disclosure, and research, interfaces effectively with Sybase, MySQL, and Red Brick databases. Also, Sybase, MySQL, and Red Brick databases operate with UNIX/Linux and Windows operating systems rather than being restricted to a single operating system, e.g., Oracle Rdb database, which NASS considered using for the 2002 Census of Agriculture, only operates in a mainframe environment and Microsoft's SQL Server database only operates with the Windows operating system.

Some information technology (IT) experts declared Red Brick a dead product in IT magazines over a decade ago when Red Brick was purchased by another company. These IT experts predicted incorrectly -- Red Brick is alive and well today. Similarly, some IT experts question the long-term sustainability of Sybase and MySQL, and therefore often tend to recommend Oracle or Microsoft SQL Server as enterprise database products. Some organizations have taken their advice in the past, and later regretted it. Predicting the future of IBM/Red Brick, Oracle/MySQL, and SAP/Sybase over the next five years, for example, is speculation. Therefore,

NASS will continue to use our existing proven database products for this transformational initiative since they satisfy the performance needs of our survey and census data processing and are economical. For example, why walk away from the outstanding performance of our 7 billion record IBM/Red Brick Data Warehouse when the product is still supported by IBM, is meeting our business needs, is easy to administer, and is economical? Migrations to other database products can be challenging and should only be pursued if there is a significant business value advantage for the Agency to change, which there is not at this time.

Some of the NASS database designs, such as the Question Repository and Quick Stats, should need only minor modifications based on additional business requirements discovered when migrating applications to centralized databases. Some databases are expected to need additional tables, such as the Metadata Repository. Some new databases will be designed, such as a Survey Management database. A description of each of the databases follows. In addition, Figure 4 provides a graphical depiction of the NASS enterprise databases. The databases shown in *italics* are the OLAP databases.

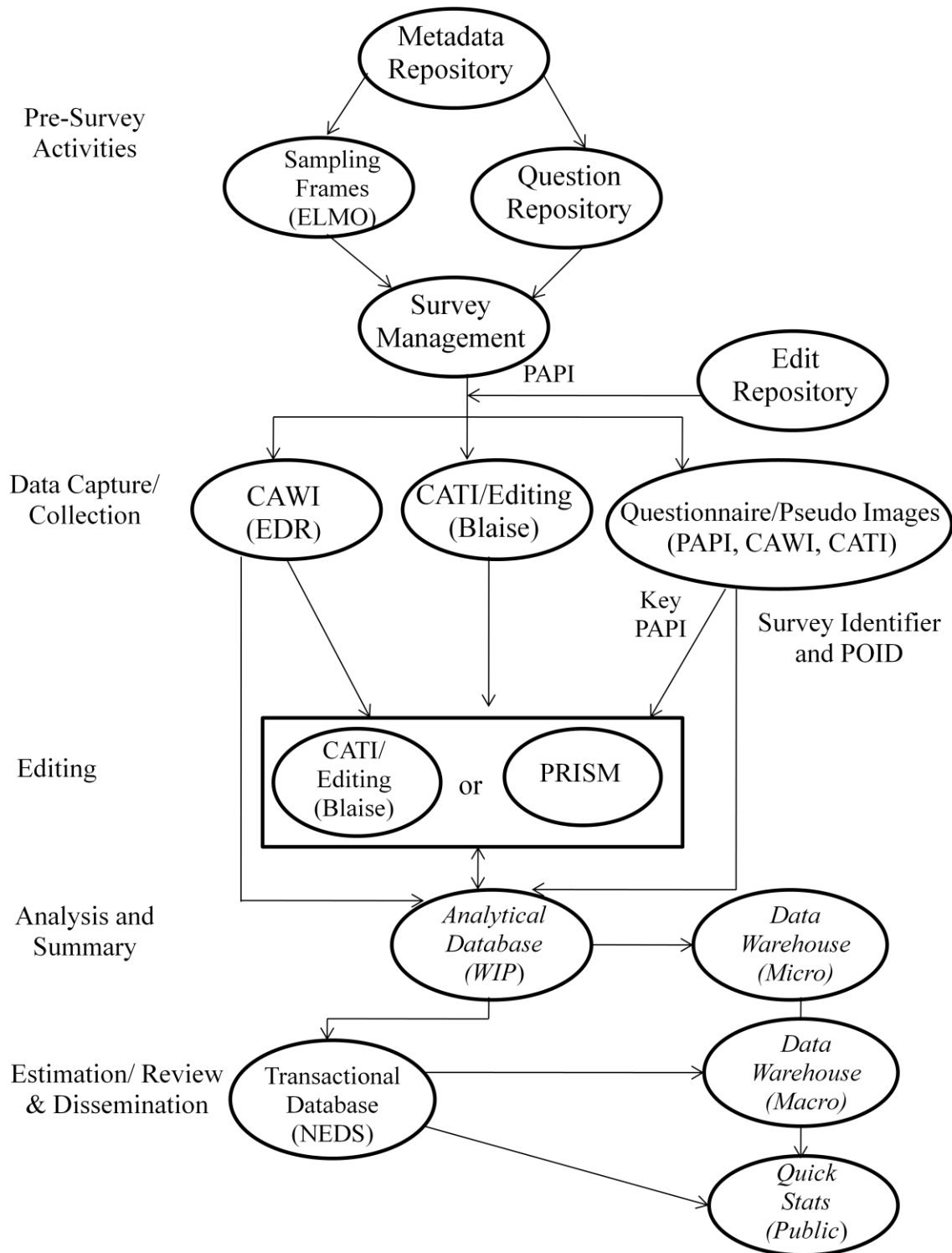
(1) Pre-Survey Activities

The *Metadata Repository database* has already been developed and is being used to provide standard metadata descriptions (variable names and item codes) to the Question Repository database and aggregate metadata to our aggregate applications, such as the Publication Tool (called PubTools). The Metadata Repository will be modified based on business requirements introduced as we reengineer or retool applications. Examples of metadata for an individual farm are variable name, item code, unit of measure, survey source, commodity, and year. Examples of metadata for aggregates are estimator type, release date, and non-disclosure/confidentiality indicator.

The *Question Repository database* will contain all survey and census questions used for NASS and State reimbursable surveys. It serves as the source for the design and building of survey questionnaire instruments. That is, the Question Repository database will serve as the source for survey questions 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). It is an internal database and the contents are not confidential. 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 is a confidential database with personally identifiable information (called PII) so is for internal use only. 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). The Record Linkage database is a repository for new list sources before new names are potentially added to ELMO and is also part of the Sampling Frames Database.

Figure 4: NASS Enterprise Databases



Changes will be made to ELMO to support the list frame activities in the National Operations Center in St. Louis, Missouri. For example, the Frames Maintenance Group that will be part of the National Operations Center will be responsible for reviewing and processing frame updates for all Field Offices. Field Office staff will be able to interactively request list frame updates and new operation additions to ELMO. To handle this new operating model, tables will be added or augmented in ELMO or a separate centralized database will be created to support these activities (known as Enhanced List Maintenance Assistant or ELMA).

The *Survey Management database* is a new database, and will focus on survey preparation and management. 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(s), and telephone interviewer availability and skills for setting up telephone interviews. This database will contain information to facilitate survey coordination across surveys and is for internal use.

The *Edit Repository database* will store edit logic that will ensure data consistency and edit limits (upper and/or lower limits) that will ensure data reasonableness. This database is under development. 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, such as the PRISM edit. It is an internal database and the contents are not confidential.

(2) Data Capture/Collection

Data collected and edited using the Computer-Assisted Telephone Interviewing and Interactive Editing (CATI/IE) 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 using the *EDR* (Electronic Data Reporting) application is already captured in a *centralized database*. The EDR centralized database is currently used for CAWI in 252 surveys and for CAPI testing in several Field Offices. The EDR database was also used to capture data from almost 100,000 respondents during the 2007 Census of Agriculture.

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. This database stores confidential individual reports for internal use and serves as a reference source for survey operations.

(3) 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 are for internal use and will store the final edited data.

(4) Analysis and Summary

The *analytical database, called WIP (Work In Progress)*, will be a single OLAP database for all census and survey data and will contain multiple years of data in addition to the current survey's data. This database will provide the current and historical data for analysis and summary processes, such as the Interactive Data Analysis System (IDAS), the PRISM analytical tools, and the Survey Processing System (SPS) Summary. The WIP analytical database will also contain the information needed to administer or 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 database will continue to contain the deep history of farm, ranch, and agri-business data and currently contains survey and census data from 1997 through 2010. It is an OLAP database for internal use only.

(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. Information includes survey indications, Field Office recommendations and comments, and the approved Agricultural Statistics Board (ASB) estimates. This transactional database also provides an audit trail of ASB actions and official revisions. It contains the current aggregate information as well as historical aggregates, e.g., past ten years, used in the estimation/review process. This database has already been designed.

The historical macro-level or aggregate information will be loaded to a *Data Warehouse* which will be the analytical database containing all historical indications, estimates, et cetera for additional analysis and research purposes. This database, also called the *NEDS Analytical database*, has been designed and is already partially populated with historical information. Two loading efforts are required. The first effort is to load aggregate information that currently resides in the NEDS FoxPro tables. The second effort is to load, for the first time, information from other sources that are being used or need to be used in the estimation or review process. This involves loading survey indications, both current and past, and a considerable amount of

administrative data. Any information that is used in the estimation/review process needs to reside in this database.

The *Quick Stats database* was recently deployed for public use. The previous Quick Stats database consisted of over 250 disjoint database tables lacking referential integrity, which created an environment prone to data and metadata inconsistencies. The new Quick Stats database integrated these disjoint tables into 5 integrated database tables with consistent metadata. About 10,000 data errors or inconsistencies have already been identified that have existed in Quick Stats historically and about half of the data inconsistencies have been resolved so far. NASS should commit to resolving as many of the remaining known data inconsistencies, as possible. This database includes only the official estimates that are released to the public, and the goal is for the Quick Stats database to serve as the data source for all public data products.

Finally, *Administrative databases* will also be utilized to support the internal operations of NASS, such as personnel, budget, and payroll. Some of these databases are NASS administered and some are developed and managed by others, such as the Department of Agriculture. A new administrative database will be developed to replace the *ASFE* (Administrative System for Enumerators) database to provide administrative and financial information on data collection.

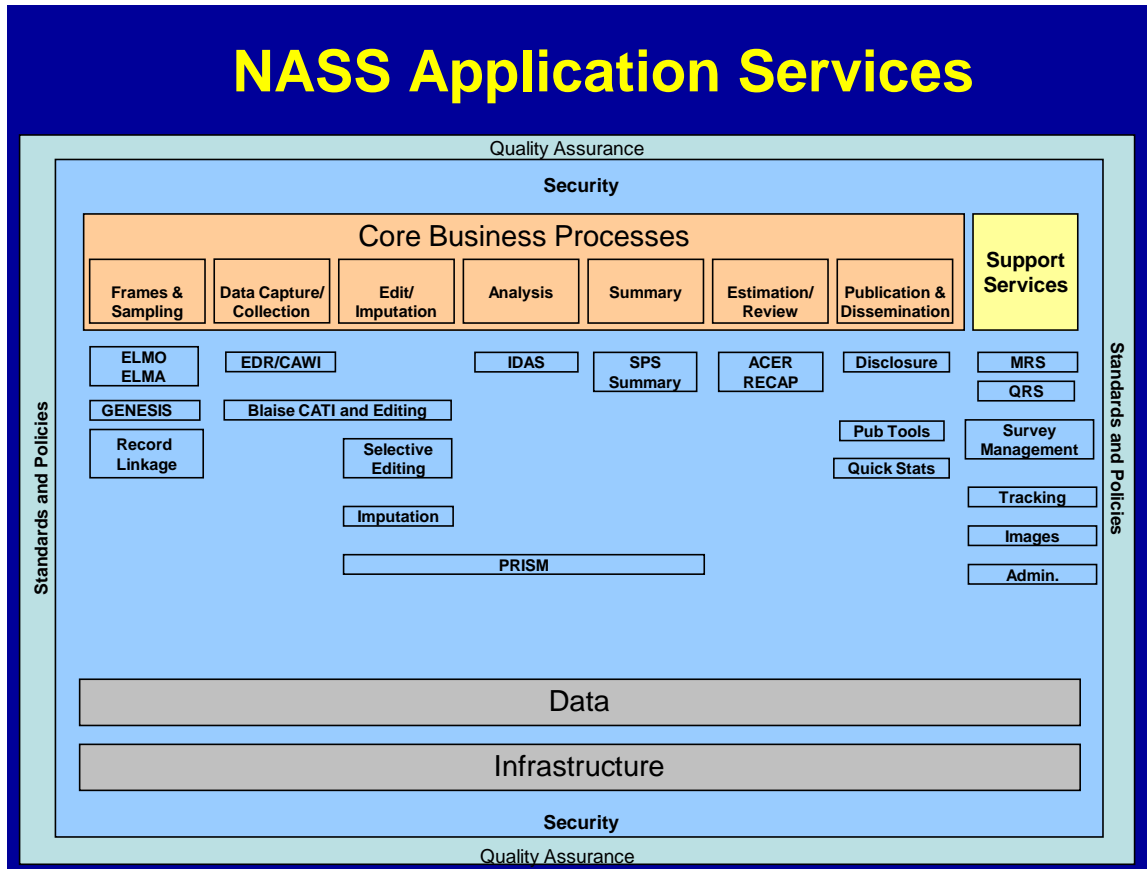
C. Application Services: Some current applications already operate in a centralized database environment, such as ELMO, EDR, and Quick Stats. Applications, such as PubTools, have already been modified to access the centralized databases. Some applications will only need minor modifications to operate effectively with centralized databases, such as the SPS Summary. Other applications, however, will need to be completely reengineered, such as the SMS (Survey Management System).

For this transformational initiative, depending upon the processing platform, we will utilize the appropriate application tools to provide the appropriate business services at an appropriate cost with an appropriate level of technical support. For example, Java, which is platform independent and is a widely used programming language, runs on both UNIX/Linux and Windows platforms and is an appropriate application framework for our survey and census data processing in our mature UNIX/Linux platform. Similarly, Microsoft .Net runs on the Windows platform, and is an appropriate application framework for our Windows services. SAS, which is also platform independent and used for sampling, editing, imputation, analysis, summary, disclosure, and research, continues to be widely used as a statistical software package on both UNIX/Linux and Windows platforms for our survey and census activities.

The plan is to provide generalized applications not only for the 153 non-standard surveys, but also for future state reimbursable surveys. A graphical user interface or GUI will eventually be developed that displays all generalized application services related to data collection, editing, imputation, analysis, summary, estimation/review, and dissemination. This graphical user interface will facilitate the use of the generalized application services and enterprise databases for all surveys.

NASS has identified seven core business processes plus support services for our survey and census program (see Figure 5). The application services will now be discussed for the support services and each of the seven core business processes (Frames & Sampling, Data Capture/Collection, Edit/Imputation, Analysis, Summary, Estimation/Review, and Publication & Dissemination). However, first, a brief discussion of our future Quality Assurance Program will be described that will encompass our core business processes and support services.

Figure 5: NASS Application Services



NASS is developing a more robust *Quality Assurance Program* to improve our survey and census programs. A quality assurance program is initially being developed for the National Operations Center to capture information to allow for the development of quality measures for critical survey activities, such as telephone calling, questionnaire or forms processing, and list frame maintenance. The use of quality measures to assess and improve our survey and census processes will eventually be extended beyond the National Operations Center to cover more survey and census processes.

(1) Support Services

There is currently a *Metadata Management Tool* that is an application providing role-based access to metadata in the Metadata Repository. This application is also referred to as the MRS (Metadata Repository System). It includes a web interface that allows for browsing metadata, editing metadata, and loading new metadata, such as Master Varnames and Master Varname Descriptions. This application will likely need modifications based on the metadata standards that are now being developed.

The *Question Repository System (QRS)* is an application that provides the questions used for PAPI, CAWI, and for CAPI testing. The plan is to use the QRS to feed questions to Blaise's computer-assisted interviewing system in the near future. The QRS will be retooled to improve code readability, simplify code maintainability, and implement any needed functionality to accommodate interaction with Blaise CATI for the 153 to-be-standardized smaller surveys.

Survey Management will be very different in a centralized database environment. The multi-purpose Survey Management System (SMS), which is currently a distributed LAN-based FoxPro system, will be replaced with a series of application services accessing centralized databases. Examples of the new application services that will replace the current SMS are: Enterprise Calendar, Data Collection Strategy, Employee Availability and Skills Inventory, Data Collection Preparation, and a Survey Management Information System.

There will also be additional support services developed primarily to support the National Operations Center in St. Louis, Missouri. These application services are: Tracking and Control, Key From Image, and Pseudo Image Creation.

The *Tracking and Control* application will replace the existing check-in features found in the Survey Management System. This application is a production control system to manage and process centralized data collection of NASS surveys. The application will be used to check in mail receipts, move work among processing units, and monitor the workflow. It will also create an instance for each critical processing step, provide an audit trail at the individual record level, control logical flows of the physical report forms, and generate management information reports. This application will work off 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. A Key From Image application is required to ensure the highest level of data quality while providing cost saving measures with increased efficiency. The Key From Image service will be flexible and allow for a mixture of heads-down data entry and key from image. The application will include user levels for data entry, verification, and adjudication.

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. NASS refers to this process as pseudo imaging.

Finally, the **Administrative System for Enumerators** (ASFE) will be reengineered to provide administrative and financial information about data collection using a centralized database.

(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: LookUp (simple fact checking on an operation), UpdateData (update data in ELMO), Extract (create output of user-selected database rows and columns), DupCheck (generate reports on matches or linked records), and Reports (create various reports, listings, and counts). Since these applications already operate effectively with a centralized database in the UNIX environment, application modifications are not expected.

The **Enhanced List Maintenance Assistant (ELMA)** is a new application that is being developed to support the centralized list frame activities from the National Operations Center. ELMA will provide an automated process to capture and review list frame update information currently presented in the List Frame Actions and Review paper reports (generated within all electronic data collection instruments) as well as list frame update information submitted by NASS Field Offices. Field Office staff will utilize ELMA to interactively request list frame updates to the sampling frames database, including name, address, telephone number, comments, and data collection mode and to request new farm/ranch operation additions to ELMO. Employees at the National Operations Center will utilize ELMA to process ELMO updates/requests on an interactive basis.

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. GENESIS also evaluates selected samples. In the future, information from GENESIS will be loaded to ELMO rather than creating many survey-specific sample master files. Other databases, such as the WIP analytical database, will then access sampling information from ELMO.

The **Record Linkage** application is used to remove duplication from list sources by identifying records believed to correspond to the same entity. Records are grouped together or matched based on the likelihood they represent the same individual. This application already operates successfully off a centralized database.

(3) Data Capture/Collection

The **Electronic Data Reporting (EDR)** application works directly with the Question Repository to create and execute web-based data collection instruments called CAWI (Computer-Assisted Web Interviewing). About 250 NASS surveys are already using CAWI, and more will be added, when practicable. The EDR application already works effectively with centralized databases.

EDR is also being used for CAPI testing in three Field Offices since EDR is thin-client, database-optimized, is already being used for web surveys with CAWI, and most NASS surveys are already in EDR for CAWI. The goal is to implement CAPI in three test states (New Jersey, Indiana, and Washington) and five additional states (Maryland/Delaware, Kansas, Florida, South Carolina, and Minnesota) by the end of 2011. The current solution not only uses EDR, but also WiFi technology and Apple iPads.

At the time of the initial testing, Blaise CAPI did not satisfy all 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. Blaise also has a new capability called CARI (Computer Audio Recorded Interviews), which saves sound files and screenshots for a sample of questions for quality assurance purposes. This would be a useful tool in the future for interviewer evaluation, data quality assessment, evaluating question wording and placement, and as a training tool. Several benefits from implementing CAPI will be improved timeliness of data collection, reduced data entry costs, and improved data quality.

NASS has been using *Blaise CATI* software since the beginning of the 1990's. Blaise is a thick client application and is distributed across our Field Offices on the Local Area Networks (LANs). We will be deploying Blaise CATI in a centralized database environment, which will be much more efficient since there will no longer be the need to create, manipulate, and transfer thousands of Blaise files. Also, there will no longer be the need to distribute LAN updates to the Field Offices when a change is made to a Blaise application. Finally, we 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 the 153 surveys that are being standardized. Therefore, we plan to retire the Electronic Data Collection (EDC) CATI/editing system that was developed within NASS.

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 or beginning of 2013. Therefore, at some time during 2013, NASS plans to assess the potential of using Blaise's CAWI offering in lieu of the EDR application.

Blaise also has added "verification" to its data entry application called CADE (Computer Assisted Data Entry) so this application should be investigated as a potential replacement for the Viking Data Entry software used at NASS now.

(4) Edit/Imputation

The Edit Repository database will store edit logic and edit limits, which will eventually be used by *EDR/CAWI, CAPI, Selective Editing, Imputation, Blaise CATI/Editing, and PRISM editing*. The PRISM edit system 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 SPS Edit will migrate to PRISM or Blaise editing so that the SPS edit will be retired. Since the EDC CATI/editing system is also being retired, there will be two edit systems retired.

The PRISM edit system has been used for the Census of Agriculture, Census of Agriculture Content Test, the National Agricultural Classification Survey, and Census Follow-on Surveys or Censuses, such as the Farm and Ranch Irrigation Survey, Organic Production Survey, On-Farm Renewable Energy Production Survey, and the Census of Horticulture Specialties. The PRISM edit system has also been used for the ARMS Phase III Survey and Cash Rents. The plan is to phase in the PRISM edit system for the following surveys: ARMS Phase II, Fruit Chemical Use, Vegetable Chemical Use, Nursery and Floriculture Chemical Use, Post Harvest Chemical Use, and the June Area Survey. The Blaise edit system will be used for other surveys.

As part of NASS's efforts to increase operational efficiencies, the Research and Development Division is researching the potential for streamlining the labor-intensive, manual edit/review process for many of our Agency's surveys. *Selective editing and automated outlier detection* distinguish responses likely to have a large impact on the survey results (which we would continue to handle through our current review and correction process) 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 statistical/selective editing and automated outlier detection approach is Banff from Statistics Canada. Banff is the most internationally recognized survey data editing application available that implements all widely-accepted statistical data editing and imputation principles. The statistical editing and outlier detection application will access the centralized database containing the data from all data collection modes. It will automatically identify potential outliers to be manually reviewed and all other records will be statistically edited and automatically corrected, when needed. The selective editing approach will then identify and flag each record whose change in its weighted survey value due to action by the edit process would have a non-trivial impact on the survey results. Subsequently, the traditional edits, such as the Blaise edits, will be run only on records flagged as outliers, those flagged for review by selective editing, and those the automated correction process could not resolve. Focusing manual review and correction only on these records will make the edit review process much more efficient. NASS will especially realize efficiency gains from selective editing and automated outlier detection when conducting surveys with large sample sizes, such as Cash Rents, County Estimates, and Quarterly Crops/Stocks Surveys.

A parameter-driven *Imputation* application is also being developed using the Banff system. This SAS application allows for numerous imputation methods while also incorporating the survey's edit parameters into the process. NASS can select from approximately twenty pre-defined imputation algorithms. Some examples of the pre-defined methods or algorithms include mean imputation and regression modeling. An assessment is underway to determine which imputation methods to use for our diverse survey program. We will also continue to provide nonresponse adjustment through the use of the reweighted and adjusted estimators in the Survey Processing System (SPS) Summary. The goal is to provide a dynamic imputation application service that will produce accurate and efficient results for all surveys.

(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, and retool IDAS to work optimally off the centralized WIP analytical database rather than countless distributed SAS files. Progress has already been made with the development of *GIDAS* and with testing in a centralized analytical database in the UNIX environment. The capability will be provided in the interactive data analysis system to interact directly with the Blaise or PRISM editing application so that data can be modified based on the analysis and then the edit application executed again to ensure data consistency.

(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 and State-sponsored surveys. Our multiple estimators are already coded, debugged, and serviceable in the SPS Summary. NASS employees are familiar with the output tables and diagnostic statistics produced. The SPS Summary will also work off the WIP analytical database in the future. We will capitalize on the proven strengths of the SPS Summary and develop tools which make it easy and efficient to create summaries for the 153 non-standard surveys and State-reimbursable surveys. The SPS Summary will then feed the survey indications 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 LAN environment with FoxPro tables for this estimation process. This process is not efficient and has been shown to be prone to data errors. The NASS Estimation application in the future (still being called *ACER*) will replace Formula1 and Lotus123 spreadsheets with generalized web spreadsheets that are integrated with the NEDS centralized database. The goal at NASS is to eventually publish survey and model-based estimates as much as possible, as the methodology and systems become available, so NASS provides more transparency and repeatability to the process than the current estimation process.

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 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 will be developed for accessing commodities or items for review or estimation from RECAP or ACER.

The *Database Integrated County Estimates (DICE)* application supports estimation and review for the National County Estimates Program. Crop and livestock county estimates, in addition to Cash Rents county estimates, will process through DICE in 2011. The DICE application enforces rounding, suppression, and publication standards, and provides for consistency in the review process. This application was originally designed to be a transitional application until county level estimation functionality is made available in ACER or RECAP. The DICE database is modeled after the NEDS database, with modifications that allow for a transition from the legacy system. The plan is to eventually use the new ACER or RECAP application and the NEDS transactional database for DICE.

A SAS-based application is used to run *Disclosure* on the Census of Agriculture and all Census Follow-on Surveys. 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 complimentary 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. The Disclosure application has also been used for the Floriculture Survey to identify primary suppressions, and should be leveraged for more surveys, where applicable.

(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 needs to be improved for this application, such as having PubTools accessed through a simplified graphical user interface, and enhancing the documentation and tutorials.

The new *Quick Stats* application was implemented recently that leverages the new integrated Quick Stats database. The new Quick Stats application provides: “Build Your Own” query capability, “Data By Subject or Commodity” query capability, “Keyword Search” query capability, and “Pre-Defined” (or Pre-Designed or Canned) queries. The new 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. During the past three months, data users have executed over 169,000 queries using the new Quick Stats application and database.

NASS subject-matter specialists need to build many commonly-requested queries for data users on our web site under the Pre-Defined queries. 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. A User’s Guide is available from the Help button in the Quick Stats application to help data users become familiar with the new application. In addition, two 4 to 5-minute videos have been developed on using the new Quick Stats application and these videos will be available to data users very soon.

V. Critical Success Factors: In the October 21, 2010 edition of Federal Computer Week, it says: *“Asking why major systems fail with regularity is like asking why there are no more .400 hitters in Major League Baseball. Success is simply too difficult.”* With this in mind, fifteen critical success factors have been identified that will guide the efforts on this transformational initiative. A brief description of each follows.

- (1) **Embrace Our Operating Model:** Our model is to provide cost savings and data quality improvements by transforming to a more flexible, more integrated, more streamlined, and more standardized processing environment throughout NASS.
- (2) **Business/IT Partnership:** We must maintain an effective business and IT partnership throughout each project – from the definition of the scope of the project and gathering of the business specifications to the application design, development, testing, implementation, and maintenance phases.
- (3) **Phased Not Big Bang Approach:** Big projects are inherently risky, complicated, and have high failure rates. This initiative will be divided into simpler, smaller projects or phases and prioritized based on expected Agency benefits.
- (4) **Achievable Deadlines:** We will avoid defining very tight timelines that are dependent on heroic staff efforts. We will have realistic expectations based on budget resources and the availability of staff, given other operational work demands.
- (5) **Clearly Defined Priorities:** We will prioritize and focus on the efforts that will provide significant cost savings and data quality improvements.
- (6) **Well-Understood Business Requirements:** It is estimated that about 70 percent of information technology projects fail before a line of code is written, due primarily to improper requirements gathering. Therefore, thorough and unambiguous business requirements need to be obtained.
- (7) **Proper Use of Contractors:** We will determine where we can benefit most from contractor assistance, e.g., gathering business requirements, designing the technical solution, conducting independent technical assessments, and technical writing. Contractors will be provided the proper architectural and technical framework and standards to follow so the products developed will be readily supported and maintained by NASS employees. Also, NASS employees will shadow application development efforts by contractors to ensure a smooth transition.
- (8) **Project Management:** We need to achieve our objectives within the defined time schedule and cost constraints so we will manage our projects effectively using sound project management principles.
- (9) **Team Collaboration:** A critically important role of the project leader is managing the soft people issues so we cultivate effective work relationships among the team members.

- (10) **Continuous Communications:** We will provide frequent communications to all employees. A monthly status report will continue to be prepared and distributed. The 15-5 rule will be used (the update should take no longer than 15 minutes to write and take the reader no more than 5 minutes to read).
- (11) **Generalized Services:** We will develop generalized application services for processing more efficiently the 153 surveys that are currently not standardized across our Field Offices, but will be standardized.
- (12) **No Exceptions:** In searching for cost savings and data quality improvements, all parts of the processing environment are open for review, for enhancements, for reengineering, or for elimination.
- (13) **Thorough Testing:** Applications will not be implemented until they have been thoroughly tested in Headquarters and selected Field Offices to ensure they address not only the functional requirements, but also the non-functional requirements, such as scalability and performance.
- (14) **Post-Implementation Reviews:** We will document what we want to accomplish with each project and then measure if we achieved the intended outcome. We will capture what went well and what did not go well so that the lessons learned become a knowledge repository for future projects.
- (15) **Continuous Improvements:** We need to follow the NASS System Development Life Cycle (SDLC) standards and continue to incorporate valuable enhancements in our application services.

VI. Implementation Timeline: As stated in Critical Success Factor (3), we will take an incremental approach to this initiative where most of the transformation will be completed during the next four years. The following quote, by Michael Petrisko in the 2010 edition of SMART Enterprise, 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.”* Also, NASS Decision Memorandum 4-10, which was approved on August 25, 2010, called for up to a five-year time period for the standardization, application development, testing, and deployment for all non-standard surveys. Our goal is to accomplish most of the activities within the next four years.

The various projects to be completed during each of the next four years will now be presented.

Phase 1: Completed By September 2011

1. Blaise is in a centralized database for half the current Blaise surveys
2. Blaise CATI/Editing is interfaced with the Question Repository and Edit Repository
3. Determine enhancements needed for the Blaise telephone call scheduler to make use of external information, such as calling history on sampled operators

4. SMS is enhanced in Citrix to satisfy the needs of the NOC
5. Develop the Tracking and Control System for the NOC
6. Enhanced List Maintenance Assistant (ELMA) is developed, tested, and ready for use
7. Enterprise Calendar is developed for initial use at the NOC
8. Employee Availability & Skills Inventory database is available for the NOC
9. Metadata: The business and technical requirements are defined
10. NEDS database is populated with 300 of 500 FoxPro files plus other files
11. WIP and MRS databases are updated for new business needs
12. Any other critical survey management functionality needs are documented for the NOC
13. The ASFE specifications are gathered
14. Multi-Level edit capability is developed in PRISM
15. Automated statistical/selective editing is tested in research and operational environments
16. Generalized imputation modules are available for use
17. Generalized IDAS (GIDAS) and Summary are developed and in production for some surveys
18. RECAP is deployed for some commodities (Review Estimates & Comments, Approve & Publish)
19. Lotus123/FoxPro is re-engineered for many commodities
20. ACER/Formula1/FoxPro is re-engineered for numerous commodities
21. DICE is utilized for county estimates for cash rents, crops, and livestock
22. Pub Tools works off NEDS and Quick Stats for some reports
23. Quick Stats: Transition to new Quick Stats products through the deployment of many pre-defined queries or tables for data users

Phase 2: Completed By September 2012

1. Blaise is in a centralized database for most current Blaise surveys
2. Blaise CATI/Editing is implemented for a number of non-standard surveys
3. Enhancements to the Blaise telephone call scheduler are implemented
4. Tracking and Control System is enhanced for the NOC
5. Data Collection Strategy & Data Collection Preparation services are available
6. Reimbursable Survey Requirements service is ready
7. Metadata: Continue the implementation of metadata standards and the refinement of metadata tools
8. NEDS database is populated with 400 of 500 FoxPro files plus other files
9. Any other critical survey management functionality needs are implemented for the NOC
10. The ASFE database is designed and implemented
11. Multi-Level edit capability is implemented for two surveys
12. Automated statistical/selective editing is implemented for some national surveys that have transitioned to the centralized database environment, such as some Hog and Crops/Stocks Surveys
13. GIDAS and Summary: Continue transitioning surveys without standard analysis and summary tools
14. RECAP is implemented for many more commodities
15. Lotus123/FoxPro is re-engineered for all commodities
16. ACER/Formula1/FoxPro is re-engineered for many more commodities

17. DICE is migrated to the centralized NEDS database
18. Pub Tools works off NEDS and Quick Stats for many more reports. A new graphical user interface is deployed for Pub Tools to simplify usage throughout our Field Offices.
19. Quick Stats: Continue developing new products

Phase 3: Completed By September 2013

1. Blaise is in a centralized database for all current Blaise surveys
2. Blaise CATI/Editing is implemented for many more non-standard surveys
3. Survey Management Information System is developed
4. Metadata: Continue the implementation of metadata standards
5. NEDS database is populated with almost all existing files
6. Multi-Level edit capability is fully implemented
7. Automated statistical/selective editing is expanded to other states and surveys, as appropriate, such as Cash Rents and County Estimates Surveys
8. GIDAS and Summary: Continue transitioning surveys without standard analysis and summary tools
9. RECAP is implemented for most commodities
10. ACER/Formula1/FoxPro is re-engineered for most or all commodities
11. DICE is replaced by ACER or RECAP for cash rents and livestock
12. Pub Tools works off NEDS and Quick Stats for most reports
13. Quick Stats: Continue developing new products

Phase 4: Completed By September 2014

1. Blaise CATI/Editing is implemented for almost all the 153 non-standard surveys
2. Metadata: Continue the implementation of metadata standards
3. Automated statistical/selective editing is expanded to more surveys, as appropriate
4. GIDAS and Summary: Transition remaining surveys without standard analysis and summary tools
5. RECAP is implemented for all commodities
6. DICE is replaced by ACER or RECAP for crops
7. Quick Stats: Continue developing new products

The NASS One-Page Project Manager (OPPM) and the newly developed project management information system will be used to track the progress of each project. The OPPM's for all projects will be posted in SharePoint early in 2011 for all employees to access.

VII. Conclusion: As depicted in Figure 6, this transformational initiative to Database-Optimized, Generalized, and Modular Applications (DOGMA) will provide cost savings and data quality gains that will improve our Agency's products and services. How significant these savings and gains 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 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 gained from this transformational initiative will position NASS to continue to provide our many valuable products and services in service to agriculture.

Figure 6: Improving NASS Products and Services

