UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Working paper
13 February 2015

Workshop on the Modernization of Statistical Production
Meeting, 15-17 April 2015

Topic (i): Business and IT changes that will impact statistical production

Information Technology Development and Implementation of a Case Management Services at a Federal Statistical Agency

Prepared by Lorna J. Drennen and Joseph L. Parsons

Lorna.Drennen@nass.usda.gov, Joe.Parsons@nass.usda.gov

United States Department of Agriculture/National Agricultural Statistics Service, U.S.A.

I. Introduction

1. NASS conducts hundreds of surveys every year and prepares reports covering virtually every aspect of agriculture in the United States. NASS's mission is to provide timely, accurate, and useful statistics in service to U.S. agriculture. Survey data collection is conducted mainly in twelve Regional Field Offices and five call centers across the United States. NASS's largest calling center is the National Operations Center that also conducts frame maintenance, training, and survey instrument development. Traditionally a decentralized organization, NASS embarked on a modernization effort in 2009. In addition to an overall restructuring of the organization and regionalization of field operations the goal was to centralize and modernize NASS's network, applications, and databases (Nealon, 2013).

2. In this paper, we will discuss the development of a centralized suite of services that are critical to the coordination and management of over 250 surveys per year. Before survey data collection, samples are selected and cases are assigned treatment codes to ensure they are available to appropriate staff and systems for correct handling during data collection. The Case Management Services (CMS) are designed to serve the following functions; The CMS leverages survey identification schemas from our metadata repository and data collection metadata from our instrument repository, the associated metadata is then used to identify and pull the selected sample for the survey from the frames repository; Once the CMS is populated with the metadata and the sample, automated events are run and when needed a manual review process is completed to prepare the cases for data collection; In addition this tool facilitates the coordination of surveys across samples, coordinating data collection treatments for cases that are in more than one survey with overlapping data collection periods; Once all cases are reviewed and a data collection treatment has been assigned the cases are flagged via an event and the appropriate mode activated in the data collection and status tracking systems.

3. NASS has several essential operating conditions that play a key role in the requirements and needs that revolve around this suite of services. The Agency has 5 National

Call Centers and over 1,000 field interviewers located throughout the Nation. The majority of NASS surveys are multi-mode and multi-location, meaning that during a survey multiple modes of data collection are run in parallel and the units surveyed can be very complex with multiple units being accounted for by one respondent. In addition, NASS has a sekew population and high burden economic surveys. In order to maintain survey cooperation, especially with large economic units that are frequently surveyed it is imperative to have an effective case manage and coordination strategy

4. The CMS is a centralized tool that consumes and delivers services to facilitate data collection and maximize survey response. The CMS services numerous layers, systems, and functions. The base functionality of the CMS is modeled off a decentralized legacy system. This new CMS comprehensively handles and or interacts with 3 of the 9 Generic Statistical Business Process Model (GSBPM) components, including design (2), build (3) and collect (4) (Appendix A). The goal of this system is to consolidate all of the activities associated with survey methodology, sampling, and data collection in a centralized location where it can deliver services to related systems and management of various functional roles. The system is robust with many automated processes and functions, reducing workloads, improving quality, and providing significantly more oversight from all levels of the agency.

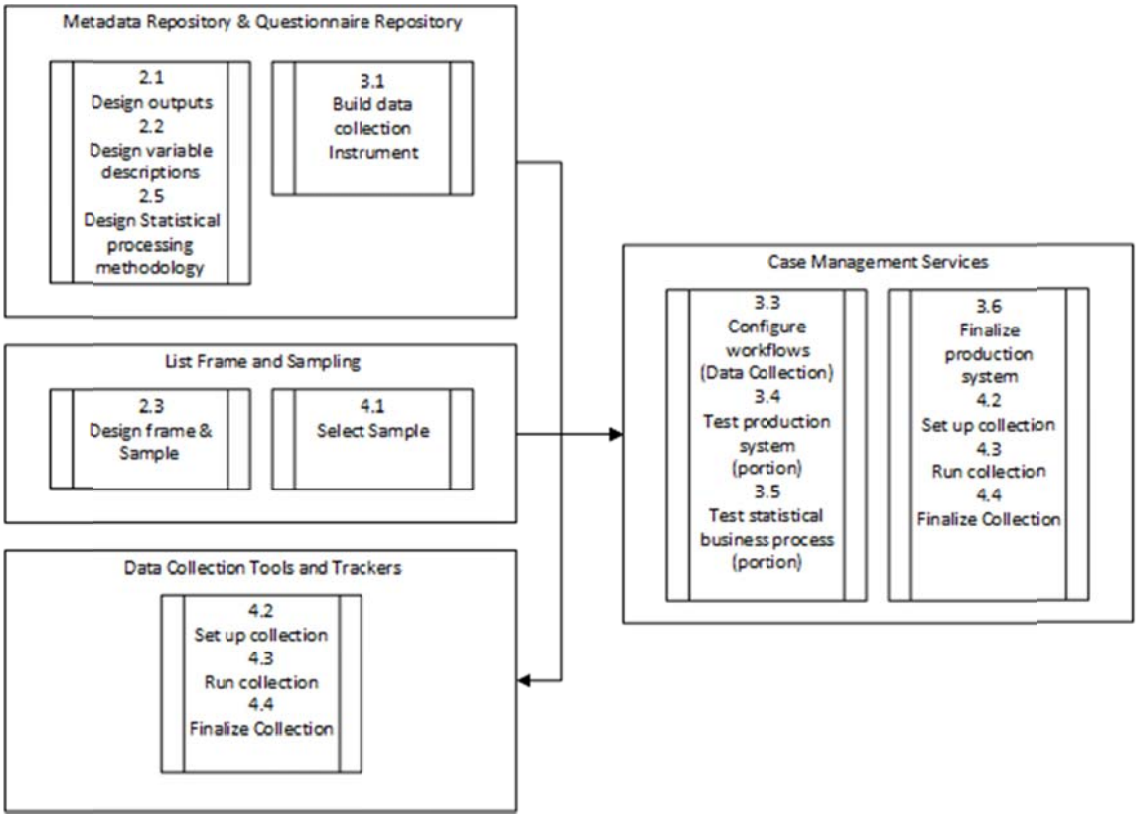


Figure 1: NASS CMS relationship to GSBPM

5. Beginning in 2009, NASS began three architectural transformations to centralize or regionalize our survey operations, primarily to provide savings in staff resource costs and to improve the quality of our statistical products. The three transformational initiatives involved: (1) centralizing and consolidating network services from 48 locations to one (Parsons and Gleaton, 2011); (2) standardizing survey metadata and integrating survey data into easily accessible databases across all surveys; and (3) consolidating and generalizing

survey operation applications for the agency's diverse survey program (Nealon, 2013). The second and third initiatives were catalysts to the success and capabilities available in the CMS. The suite of services would not be possible without their success. The addition of critical metadata, standards, and centralized databases put NASS in a position to develop against a service oriented architecture that has facilitated much of the functionality that the CMS provides.

II. Role-based Functionality

6. One of the most important elements in the design of this system is its ability to interact with different functional units and roles within the organization, allowing full systems access where needed but read only functionality where appropriate. In addition, the CMS has the ability to leverage metadata and manage survey data collection for the whole organization. This tool has given NASS the ability to strategically plan data collection at a National level. Prior to efforts to centralize survey systems and operations, data collection was managed individually in each of 46 separate field offices and 5 separate call centers. This required samples selected at the National level to be transferred to the local office for management during data collection. There was no ability to monitor data collection in progress outside the individual field office or at the National level. In addition, cases that needed to be managed by different field offices had to be transferred from one to the other. This added extra sample handling and inefficiencies.

7. The new centralized CMS has provided a number of efficiency gains in the preparation and management of our survey data collection process. Prior to the implementation of the CMS the preparation for each state included; manual set up survey instance, state specific download of sample and associated sample information, sample file processing, uploading of the processed sample file to a decentralized management tool, uploading any ad hoc comments necessary for the sample, assignment of interviews to cases, and assignment of data collection modes to cases. The CMS takes these 8 distinct decentralized processes that were being executed at a state level (up to 46 times) for each survey and consolidates them into one centralized process that is automatically executed at a National level. A conservative estimate of hours saved per year is approximately 9 full time employees.

8. The CMS is a role based tool and was designed to handle data manipulation with roles at the most granular level. At the highest level NASS has 3 functional roles with sub-roles that are currently leveraged in the CMS, including a National, Regional and Call Center Role. The National Role as allows all of the functions that were once completed multiple times by multiple individuals in multiple states to be managed at a National level by survey project managers in one location. This provides a great deal of oversight, improves communication, expedites actions, and ensure similar treatment across regions. The survey managers have the ability to perform not only high level task, but can also manage the entire survey from start to finish at a National level if needed or required. In addition the role functionally is versatile enough that a National survey manager can be located in a region and still be give the ability to run a survey as a National survey manager from their region.

9. Each of NASS's 12 regional offices are responsible for reviewing the treatment of each record and making any necessary modification from the initial treatment that was assigned. Prior to the centralization of the data, regional offices managed each state in their region separately, causing unnecessary overhead. The CMS utilizes an external role

management system and Active Directory authentication to determine what portion of the data a user can see within a region and what functions are available for their use. This allows the regions to manage the cases for their region in one place reducing a significant amount of complexity.

10. The final related high level roll in NASS is a Call Center. Call Centers need the ability to access the CMS in a read-only capacity at a National Level. This allows them to review records assigned to them and run reports to determine upcoming workloads. Call Centers also have a unique understanding of case statuses and can often observe an oversight on records without proper contact information that can be overlooked or missed by other processes. The ability to view the CMS at a National Level quickly provides them with the information needed to communicate required modifications back to the Regional Offices.

11. The automation of previously decentralized processes and the centralization of all survey cases into one database has provided a great deal of flexibility and visibility for the first time in NASS. Roles and access rights can easily be added to accommodate different informational needs making automation and information visible to all interested parties.

III. Multi-System and Database Interaction

12. As the system stands currently, it interacts with 9 different services and the addition of at least 3 other communication points will be incorporated moving forward. These services and databases drive components 1 -4 of the GSBPM within NASS.

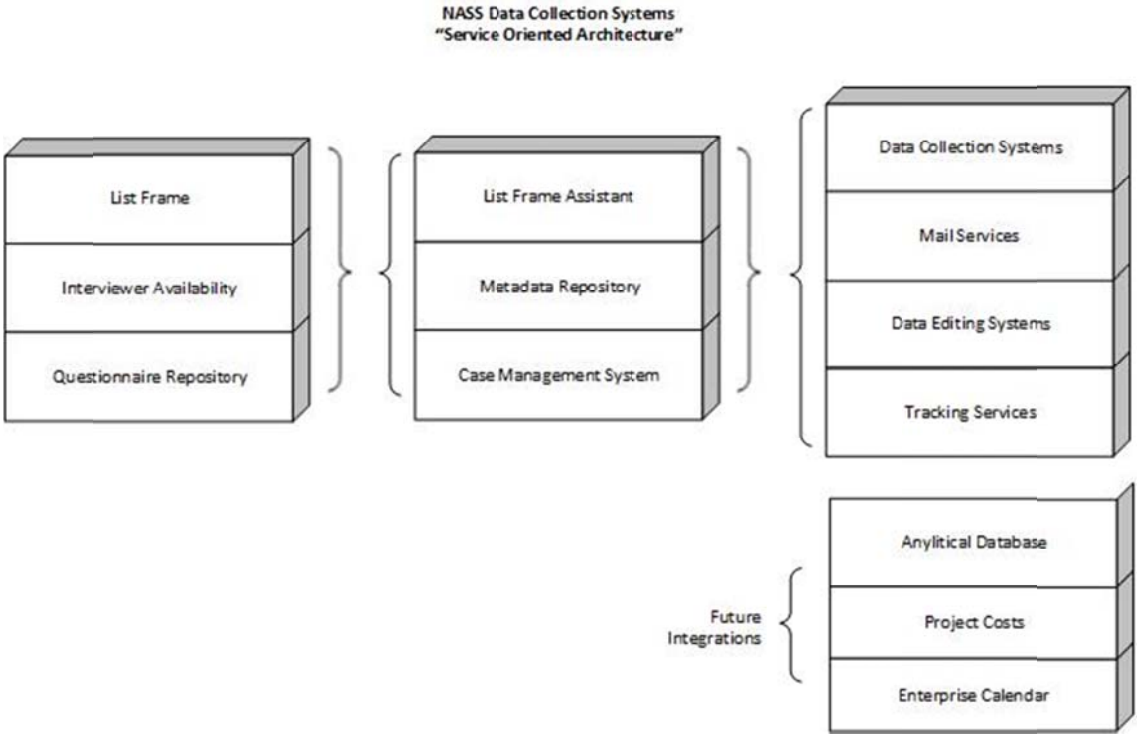


Figure 2: NASS Data Collection Systems

13. Interaction with the above services are coordinated through the CMS with "Events." Events are added to a timeline that then drives each of the process and service calls required to manage the data collection process from start to finish. Each Survey starts with a generic

set of events and the generalized events can be customized for individual survey needs. For example, one of the events is a service call to our Interviewer Availability repository which allows userto assign cases to interviewers by location who have been trained for the survey and are available to work during the data collection period. These events have given us the ability to easily leverage and automate responsive design practices. Responsive design is how we manage field work, it involves striving for efficiency and striking a balance between cost and error (Miller 2014).

14. Currently, the adaptive design principles being used through events are very simple. An example includes the use of auxiliary information for sample units such as stratum or permanent treatment codes indicating known information about individual sample unit to predetermine the treatment of a record. This treatment code then identifies a predetermined path for the handling of the record through the survey cycle. The most elaborate treatment code marks the record as eligible to be contacted in every available mode. NASS is unique to many survey organization in that 90percent of our surveys are multi-mode surveys including Mail, CAWI (Computer Assisted Web Interview), CATI (Computer Assisted Telephone Interview), and CAPI (Computer Assisted Personal Interview)/PI (Personal Interview), in addition one case can be in multiple surveys at one time. Records that are assigned a treatment code that includes all modes of data collection will be managed through a series of events. For example, if a survey has a 30 day collection widow, the event matrix may execute the following example actions:

Day of Survey	**Action**
15 days prior to start date	Mail files is created for all mail records and pushed to Mail Services
Day 1 of Survey	If not received, record is activated in CAWI and available for online reporting *Letter included with paper questionnaire will include CAWI Instructions
Day 10 of Survey	If not received, record is activate in CATI for telephone data collection
Day 20 of Survey	If not received, record is activate in CAPI for in person data collection

15. With research that takes into account total survey error and additional service connections such as NASS's analytical database and available budget allocation the above set of events could be expanded to consider the impact of the records on the complete dataset, response rates and remaining budget dollars. One way to measure the impact a records has on the complete data set is through an automated event comparing the CV of pre-summarized variables in the active dataset to the relevant control data for the record to determine if a positive report would influence the CV for those variables. The calculation of response rates and counts could prove useful in the case where a survey needs a minimum number of responses by variable by location (common for NASS small area surveys), the same action as above could be executed considering the current number of positive reports for a relevant variable in the related geographical area. Lastly, available budget is a critical component when determining if a record is to be collected in the field, use of this information in conjunction with the above examples could help right size the follow-up field sample given cost constraints using a standardized method that will determine if a missing records should remain in CATI, move to CAWI, or be deactivated.

16. Much of this is possible due to a long term investment in transactional and analytical database functionality (Nealon, 2013). Many of NASS systems have recently been modernized and leverage an enterprise transactional database. The relevant data from the transactional database then filters into an analytical database through an extract, transition, and load (ETL) process where variables are aggregated in materialized views and are available for analysis throughout the survey proper. This functionality is what will allow the Events in the CMS to retrieve and analyze data on the fly to determine the future of a record in a data collection cycle. This has an impact on our essential operating conditions (paragraph 3), as aggregated data analysis on the fly for responsive treatment could play a critical role in reducing any unnecessary touches and the retention of the long-term relationships the population requires.

IV. Discussion of Different Perspectives

17. As NASS works through the implementation and adoption of this tool we continue to find additional efficiencies. It is allowing NASS to complete actions in a much more automated fashion, eliminating a great deal of manual and redundant work (i.e. doing it in 46 locations rather than 12 or 1). As an organization we continue to study and work on methods and ways to manage and improve our response rates. NASS is unique in that we reach out to the same respondents multiple times a year, making it extremely important that build respondent trust and loyalty (paragraph 3), this new tool will allow us to develop more ways to be efficient, maintaining a balance between cost and error (Miller 2013). Other organizations may not have near the challenge of the repeat cases in multiple samples, however could still find numerous benefits with a tool as versatile, if only for multi-mode multi system interaction and efficiencies.

18. In addition NASS will benefit from the increase in management and oversight during data collection, as information is centralized and accessible in ways it never was before. The ability to have over 25 surveys processing at one time with case overlap across samples, ensuring that records are never active in more than one mode at one time and are being treated the same across surveys is powerful. The additional paradata generated for use over time will improve our techniques and process benefiting data collection moving forward. This system empowers NASS with the ability to plan smarter in the beginning and will lead to more dynamic and automated management of data collection in the future.

V. Technical Overview

19. Our case management tool was designed to be a service oriented to allow the software to keep up with business requirements and changes more efficiently. The tool was conceived with the notion that all of the more modern systems would eventually become smaller more functional service oriented tools that could interact with the case management tool and other tools. However the system still needs to provide fixed file formats to existing legacy applications. In comparing the new service oriented tool with the legacy case management system, legacy interacted in a hub and spoke manner. If a process needed to interact with the legacy system the legacy system was changed to provide one or two way communication with that process. As more and more process get added to the hub and spoke system it becomes increasingly difficult to maintain and update. The new system simply provides the mechanisms to allow other process to retrieve required data. The reading of the data requires authentication but once that is established for a process nothing additional needs to be added

to the system to read data from the services provided. There are no programatic changes required to get a new process interacting with the case management tool.

20. Early in the development phase it became clear that case management is very event driven - a new survey or data collection need comes online creating the need for new actions. To handle this an event timeline was established in the case management system. This allows for efficient use of time, rather than updating data manually by either going to a web site or running a script the system allows you to schedule an event. An event has at least one action and can have zero or more constraints. This notion of an event and triggering process to run has led to the expansion of the service oriented notion of the case management services. By allowing an outside process to register receiving service events can be put in place which will allow the CMS to push communication to the registered process. Although currently there are still legacy system to interact with that are not service based, over time as these are modernized the ability to plan out a data collection cycle from beginning to end and have it run without any manual intervention is possible.

21. With all this flexibility in mind, it became clear that the service model should not be limited to one format. At a lower level service serialization was provided first with some defaults: XML, JSON, JSONP and some lesser known outputs. However given the need to interact with legacy systems additional serializers/de-serializers were added DBF (Fox Pro), BCP (Sybase bulk loader) and CSV (Comma separated format). By providing these serializers/de-serializers at a low level independent of the specific service being provided the system can provide any of them for any request - all the client or requesting process needs to do is ask for the specific format. For example, the print mail center used to receive DBF files with the information for the labels of the respondents. By providing the DBF serializer we were able to provide that legacy system that legacy output. Eventually that system will be modernized and may use a more modern construct of data XML or JSON, but for now it did not need to change to integrate with the new case management tool.

22. The CMS application was developed to operate on a standard Linux server with a SUSE operating system. The application is developed in the Catalyst MVC (Model View Controller) Perl framework. The presentation layer leverages HTM5L, JavaScript, CSS, and Dojo JavaScript Libraries. RESTful services and AJAX communication techniques are leveraged for client/server communication. Standard JSON, XML and CSV data formats are leveraged to provide standard, diverse, and efficient input and output formats. CMS utilizes Object Relational Mapping (ORM) to facilitate communication between the CMS application and the MySQL CMS database and other databases such as Sybase and Red Brick.

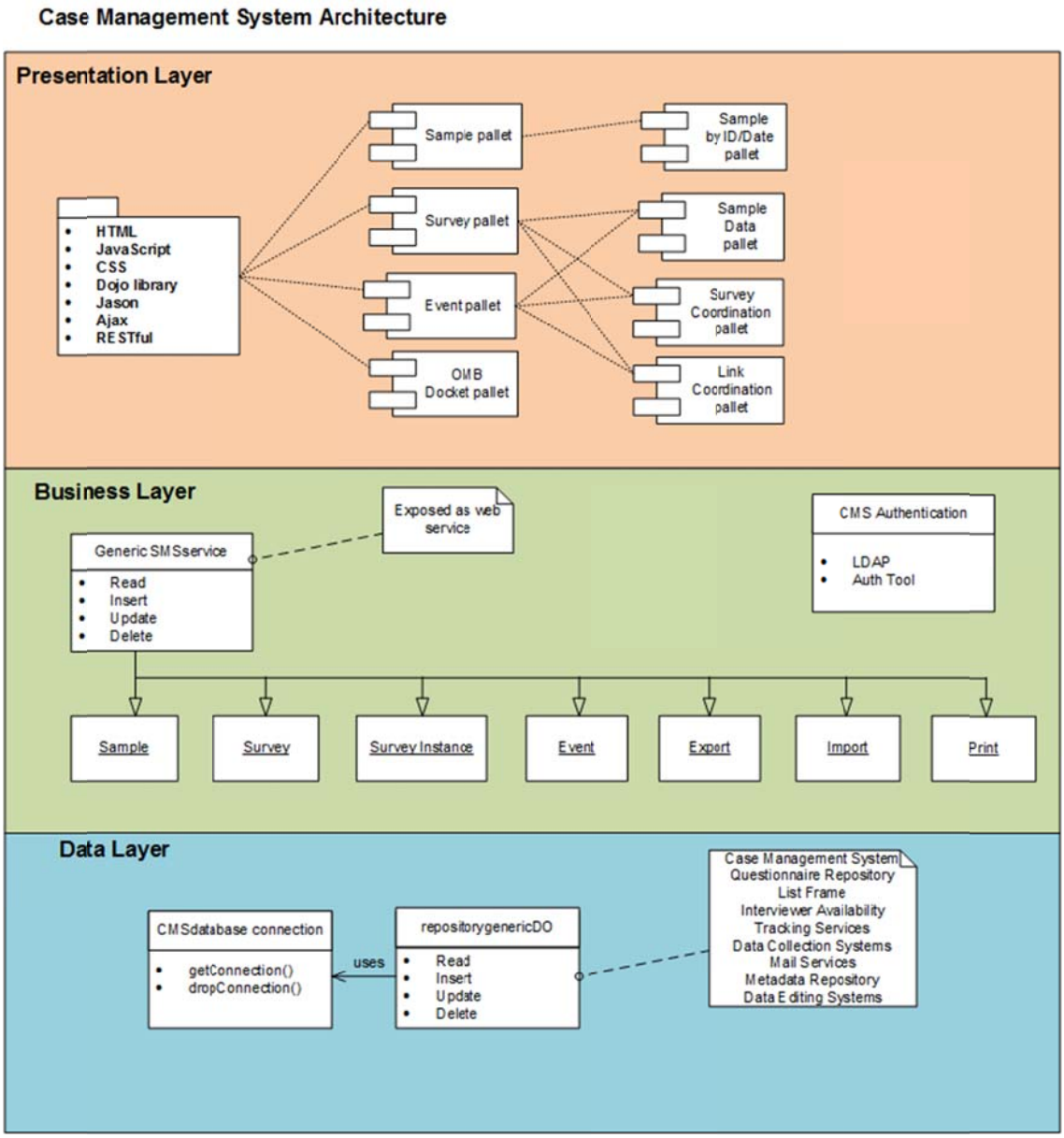


Figure 3: Case Management System Architecture

VI. References

Nealon, Jack and Gleaton, Elvera T., "Consolidation and Standardization of Survey Operations at a Decentralized Federal Statistical Agency," Journal of Official Statistics Vol. 29, No. 1, 2013, pp. 5–28, DOI: 10.2478/jos-2013-0002.

United Nations Economic Commission for Europe (UNECE), "**Generic Statistical Business Process Model**," UNECE on behalf of the international statistical community, Version 5, December 2013 http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0

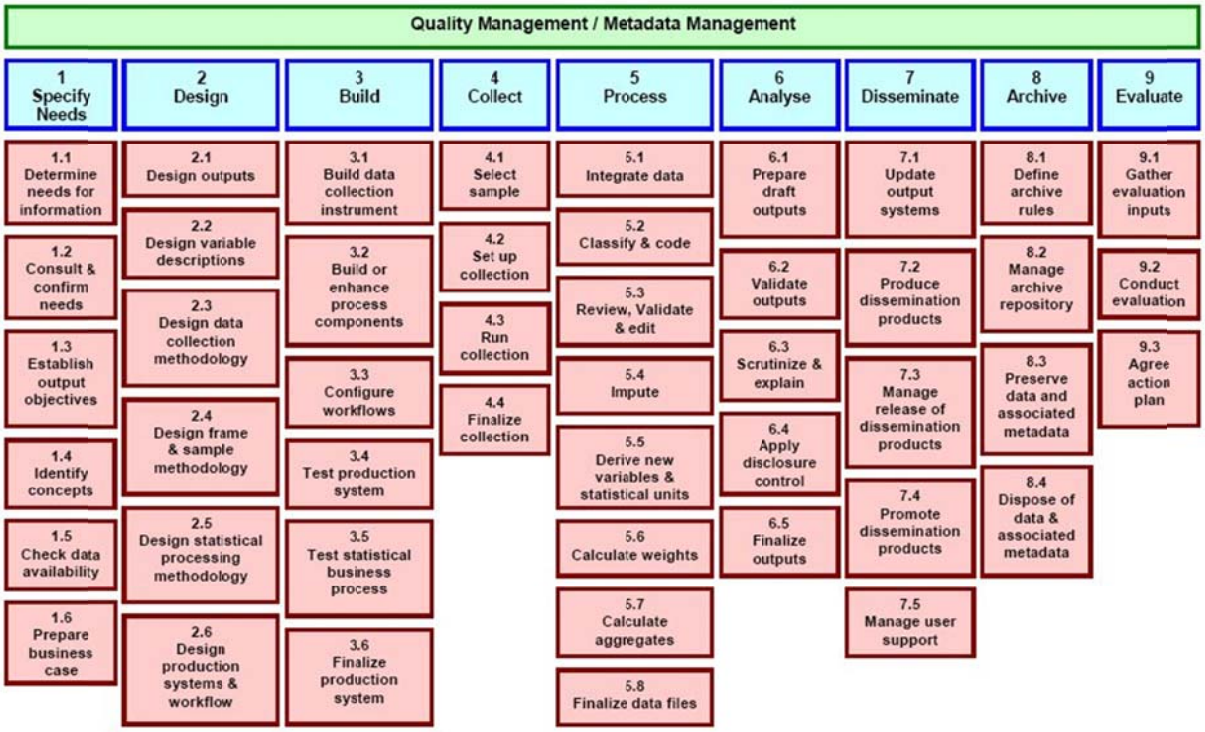
United Nations Economic Commission for Europe (UNECE). Generic Statistical Business Process Model. Accessed online February 26, 2014. http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Business+Process+Model

Parsons, Joseph L. and Gleaton, Elvera T., "Virtualizing and Centralizing Network Infrastructure at a Decentralized Federal Statistical Agency," (paper presented at the meeting on the Management of Statistical Information Systems UNECE, Washington, DC, May 21-23, 2012) accessed January 2, 2014 http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2012/06_USA.pdf.

Miller, Peter V., United States Department of Commerce, Census Bureau; "What does Adaptive Design mean to you?" (FedCASIC 2014) Accessed online February 12, 2015 https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCwQFjAA&url=https%3A%2F%2Ffedcasic.dsd.census.gov%2Ffc2014%2Fppt%2FFedCASIC%2520keynote%25203%252017%252014%2520%2520final.pptx&ei=lKXcVJ2VEcirNpfegsAD&usg=AFQjCNGJLirzXBsX6qtnm5jNB5slnem_4w

Parsons, Joseph L. and Duxbury, Brandon W., "Information Technology Centralization and Modernization Efforts and the Impact on Organizational Culture at a Federal Statistical Agency," (paper presented at the meeting on the Management of Statistical Information Systems UNECE, Dublin, Ireland April 2014) accessed February 10, 2015 http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2014/Topic_1_USA_Parsons.pdf

APPENDIX A: Generic Statistical Business Process Model (UNECE); The General Business Architecture for a Statistical Agency



http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0

The GSBPM describes and defines the set of business processes needed to produce official statistics. It provides a standard framework and harmonized terminology to help statistical organizations to modernize their statistical production processes, as well as to share methods and components.