WP.29
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (iii): Electronic data reporting – editing nearer source and multimode collections

ELECTRONIC DATA REPORTING AND DATA COLLECTION EDITS AT THE NATIONAL AGRICULTURAL STATISTICS SERVICE

Supporting Paper

Submitted by National Agricultural Statistics Service, United States Department of Agriculture[1]

I. INTRODUCTION

1. The United States Department of Agriculture's National Agricultural Statistics Service (NASS) conducts hundreds of surveys annually on the nation's farmers and agribusinesses. These data collection efforts provide the basic data from which estimates are derived for virtually every facet of agriculture production – supplies of food and fiber, economic aspects of the industry, and agricultural chemical usage. The majority of NASS' surveys are multi-modal; hence, questionnaires need to be developed for multiple data collection modes: mail, face-to-face, telephone, and most recently the World Wide Web. To efficiently create the numerous questionnaires needed, NASS developed a client-server based Question Repository System (QRS) to generate both paper and Web questionnaires. The QRS is part of NASS' electronic data reporting (EDR) system infrastructure for collecting survey data on the World Wide Web. This paper will discuss NASS' approach to EDR, including the capabilities of the QRS. The paper will include discussion on how data collection edits are built and applied on Web surveys. The implementation of these Web data collection edits will be compared and contrasted with edits utilized in NASS' other data collection modes.

II. THE NASS DATA COLLECTION PROGRAM

2. NASS' organizational structure consists of a headquarters office in Washington, DC and 46 field offices that service all 50 states and Puerto Rico. NASS' headquarters office is responsible for formulating and implementing the methodology for all national level data collection programs. This includes questionnaire development, sample design, data editing and summarization procedures, and the production of estimates. The field offices are responsible to carry out data collection for their respected areas.

3. NASS' data collection programs can be classified into two major areas: censuses and sample survey programs. The sample survey programs consist of hundreds of individual surveys that are repeated at predetermined intervals (weekly, monthly, quarterly, annually). National sample sizes range from a few dozen to several hundred thousand. The census program consists of a few censuses for specific commodities, but more notably, the quinquennial Census of Agriculture. The Census of

[1] Prepared by Daniel G. Beckler (dan_beckler@nass.usda.gov).

Agriculture attempts to contact all known agricultural operations in the United States. In 2002, Census of Agriculture questionnaires were mailed to almost 2.8 million anticipated agricultural operations. The sample surveys generate an array of the most critical, time sensitive state and national level agricultural estimates throughout each production year, while the Census of Agriculture provides a virtually exhaustive list of agricultural estimates at the national, state and county levels.

4. Most NASS data collections are multi-modal and include some combination of mail, face-to-face, and telephone. Most recently, NASS began making all of its sample surveys and censuses available on the World Wide Web in order to comply with the Government Paperwork Elimination Act (GPEA) and as a service to its respondents.

5. The data collection periods for most of NASS' major sample surveys are very short – usually only two weeks. In order to minimize non-contacts, NASS makes little use of conventional mail-out/mail-back techniques for its sample surveys. Instead, the agency relies heavily on telephone/CATI (computer assisted telephone interview) and face-to-face modes deployed simultaneously. Most sample survey data are collected via telephone, often with CATI. Data for large operations, or past refusals, often receive face-to-face interviews. Blaise software is used for CATI data collection instruments. Aside from a couple rather small research projects, NASS currently does not use any type of CAPI (computer assisted personal interview), but instead relies on paper questionnaires for face-to-face interviews. The short data collection periods generally do not permit Web versions of surveys to be made available to respondents prior to when interviewer assisted methods are utilized.

III. THE NASS ELECTRONIC DATA REPORTING SYSTEM

6. The necessity to produce Web versions of surveys also brought the need for a system to develop them. NASS' electronic data reporting system consists of the Question Repository System (QRS), a series of PERL scripts running on a Web server, and associated databases. The QRS was developed in-house and serves as the backbone of the EDR infrastructure. The vision for the QRS was to develop a system to efficiently create publication-quality paper and Web questionnaires. In order to fit into NASS' organization structure and processing environment, the QRS needed to accommodate the following requirements.

- The QRS needed to create paper and Web questionnaires without the users needing to know Hypertext Markup Language (HTML), or any other type of programming. This requirement was necessary since, due to staff resource limitations, the primary QRS users would be existing questionnaire designers who have extensive experience with word processors, but are not programmers.

- The QRS needed to store individual questionnaire components – questions, headers, footers, instructions, etc. – as individual entries in a central database repository. This requirement would allow reusing questionnaire components that are common among multiple state versions in a given survey, as well as those that are used in multiple surveys. This would provide greater standardization within and across surveys as well as increase efficiency in creating questionnaires.

- Users needed to be able to identify subsets of the individual questionnaire components in the central database to be used to generate paper and Web questionnaires simultaneously.

- The process to generate the paper and Web survey instruments needed to be as efficient as possible. This requirement was crucial since NASS conducts over 400 surveys annually, most of which have several questionnaire versions to accommodate uniqueness across states. The survey instruments for these surveys range in size from a single page with only a few questions, to over 30 pages with several hundred questions. Furthermore, due to resource limitations, all versions

of both the paper and Web questionnaires needed to be created by NASS' existing questionnaire design staff.

- The system needed to be robust enough to accommodate virtually any type of questionnaire design, yet provide tools to ensure standardization across questionnaires. Although most surveys NASS conducts are relatively static and repeated at set intervals (weekly, monthly, quarterly, or annually), there is enough variation in survey programs to necessitate a flexible system. Also, NASS routinely implements new surveys for other organizations that often collect data not encountered previously.

- The Web survey instruments needed to comply with the Government Paperwork Elimination Act (GPEA) and with Section 508 of the Americans with Disabilities Act.

- The system had to use existing Agency metadata that would benefit the questionnaire development process, as well as contribute to NASS' metadata warehouse. This would benefit other existing and future systems.

- The QRS-created Web questionnaires would need to display consistently across a wide range of browsers and computer platforms. NASS decided to use "plain" HTML to accomplish this, and not utilize Javascript, Java applets, or other "higher" Web programming languages. NASS may readdress this requirement in the future as technology continues to progress.

- The QRS needed to be developed and implemented in roughly a three-year period with a limited budget. A limited pilot QRS was developed during the first year and a production version containing essential features was developed during the second year. Third year development consisted of enhancing and refining the production system.

IV. QRS DESCRIPTION

7. The Question Repository System that was developed accomplished the requirements listed above. The NASS QRS is a client-server application consisting of three main components: (1) a user interface client written in Visual Basic.NET (VB.NET), (2) Microsoft Word with an added application written in Visual Basic for Applications (VBA), and (3) a Sybase database.

8. The VB.NET client serves as the "control panel" for the QRS. The client has three main functions: (1) it allows users to search and preview the individual questionnaire components (questions, headers, footers, etc.) as well as completed survey questionnaires, (2) it allows users to build paper and Web survey questionnaires by specifying a subset (known as a *recipe*) of the individual questionnaire components as well as supplying necessary metadata (such as the survey's title, the states the survey is valid in, the date of the survey, etc.), and (3) the client instantiates Microsoft Word (with the added VBA application layer). Microsoft Word is used to produce all individual questionnaire components.

9. Perhaps the most unique component of the QRS is the Visual Basic for Applications application that was written for Microsoft Word. This VBA application adds considerable functionality to the off-the-shelf version of Microsoft Word. Specifically, the VBA application provides a host of formatting tools that ensure standardization. Such things as font sizes, line spacing, margins, tab sets, table row/column sizes and answer box sizes are controlled by a series of custom tool bars added by the VBA application. In addition to ensuring a standard appearance for the final questionnaires, these features provide tremendous timesavings for the questionnaire designers.

10. The VBA application also provides a formatting wizard for HTML input fields (i.e., text boxes, radio buttons, check boxes, and drop-down boxes). This feature allows a questionnaire designer with literally no HTML programming experience to produce virtually any type of Web input field.

Additionally, the VBA application provides a "fill variable" input wizard that allows questionnaire designers to insert variable fields that will be replaced with current information when the questionnaire components are actually used in a survey. For example, if the current year is desired to appear in a question (e.g., How many acres of corn did you plant in 2004?) the questionnaire designer would not actually enter "2004," but rather insert a fill variable (e.g., How many acres of corn did you plant in CURRENT_YEAR). This field (CURRENT_YEAR) would be replaced with the appropriate year when the question is actually used in a survey instrument. This feature allows for greater reuse of questionnaire components, and consequently, more efficient questionnaire creation.

11. The same concept of "fill variables" is used for question numbering. The VBA MS Word application allows questionnaire designers to label questions with dynamic question numbers. When questionnaires are built, the QRS sequentially numbers all questions correctly. This feature allows questions to be able to be used in different locations in multiple questionnaires.

12. The VBA MS Word application also allows question designers to build skip patterns within questionnaires. For the designer this process involves identifying the screening question and the questions that the respondent will be routed to when certain conditions are met. This process is accomplished with a dialog box. This feature allows questions to be built with skip logic by QRS users with no programming background.

13. Finally, the VBA application provides a tie to NASS' data warehouse – specifically, the master variable names and associated metadata. This data warehouse contains a repository of all official master variable names for all of NASS' questionnaire items, as well as all previously reported survey data. The master variable names are used by a variety of post-data collection editing and summarization programs. The MS Word VBA application allows questionnaire designers to embed (as bookmarks) master variable names in questionnaire answer cells that allow Web data to be captured to appropriate variables.

14. All questionnaire components as well as all completed questionnaires are stored in a Sybase database as binary large objects (BLOBs). The individual questionnaire components are stored in both their native Microsoft Word format files (.doc files) and in MS Word generated XML format files. The MS Word files of the individual components are used to assemble completed paper questionnaires, which also are stored as MS Word format files. A Web server uses the XML files for the individual components to dynamically create Web questionnaires. In addition to the BLOBs, the database also stores many metadata fields related to the components and completed questionnaires.

15. Storing the individual questionnaire components in their native MS Word format, and using those Word files to assemble paper questionnaires provides the benefit of retaining all formatting (font, alignment, spaces, etc.) details of each component. This ensures the constructed paper questionnaires will be of publication quality.

V. OTHER QRS FEATURES

16. In addition to the many formatting related features, the QRS provides several features to aid in the development and management of questionnaires. One of the most powerful QRS features is the notion of a *recipe*. A recipe is simply a subset of questionnaire components (i.e., a header containing branding, all of the questions, a footer containing the required OMB burden statement, instructions, etc.) that are included in a questionnaire for a particular survey. These recipes are saved in the QRS database and may be retrieved at a future time to generate questionnaires for the same survey for another reference date.

17. Another useful feature is the ability to build questionnaire components for specific modes (currently only paper and Web). While being built, components are identified to be used for the Web, for paper, or for both. If separate paper and Web versions exist for a specific component, these components may be linked together to maintain their "equivalence" relationship. After questionnaires are built using

these components, the appropriate one will be used for each mode (i.e., the paper version will be used to generate the paper questionnaire and the Web version will be used to generate the Web questionnaire). Furthermore, the system allows individual portions of a component to be masked for a specific mode. This allows a single component to be built for both paper and the Web, but still allow flexibility for each mode. For example, a questionnaire instruction may request respondents to "Return the completed questionnaire in the enclosed postage paid envelope." Clearly, this line is not appropriate for the Web mode, consequently, this line could be flagged to appear on the paper questionnaire only and be masked on the Web.

18. The QRS also contains a feature to track the history of questionnaires. Although somewhat uncommon, changes to questions from one implementation of a survey to another do occur. The QRS provides users with a history of all changes to questions as well as an assurance that the most current question versions are used.

19. Another QRS feature allows questionnaire designers to automatically e-mail paper versions of questionnaires (i.e., the MS Word files) to NASS field offices or other NASS staff. This feature is used to distribute draft versions of questionnaires for review, or to send finalized versions to NASS' field offices for data collection.

20. While being primarily built to produce survey questionnaires, the QRS may also be used to produce forms used internally at NASS, which may then be delivered to NASS employees via NASS' Intranet. These forms may be for administrative data collection efforts (such as leave permission slips and supply orders) or for internal organization climate surveys.

VI. IMPORTANCE OF METADATA

21. NASS is becoming increasingly aware of the benefits of metadata in virtually all aspects of the survey process. Consequently, metadata are playing a key role in the development and operation of the Question Repository System. Metadata generated by the QRS and by other systems are used extensively by the QRS to ensure accuracy in the generated questionnaires, as well as to allow the QRS and other NASS systems to interact more efficiently. Notable examples of the uses of metadata by the QRS include:

- All survey components are identified as to which states they apply. This information is used to ensure that the appropriate questions are delivered to the appropriate states. For example, NASS conducts a series of Quarterly Crops/Stocks Surveys each year that collect acreage and production information for a series of crops. The specific list of crops the surveys target vary by state. By attaching state metadata to each question, QRS users are able to create individual state questionnaires that contain the correct list of crops by building only a single recipe for the survey.

- The master variables obtained from NASS' data warehouse contain metadata for such things as a variable description, the type of data the variable would store (text, numeric, dollars/cents, etc.) and a measure of precision. All of this information is used to render the Web input field correctly, as well as to provide HTML alt tags for Section 508 compliancy.

- The questions the QRS produces serve as metadata for other systems. Hence, a user retrieving previously reported survey data from NASS' data warehouse can not only retrieve the survey answers, but also the question text that was used to collect the data.

VII. NASS EDR ARCHITECTURE

22. The Question Repository System, along with its associated database, is the heart of NASS' overall questionnaire production environment (except for CATI instruments). However, it is only one part of NASS' complete electronic data reporting (EDR) system. Figure 1 illustrates the entire NASS EDR system for generating and delivering Web questionnaires.

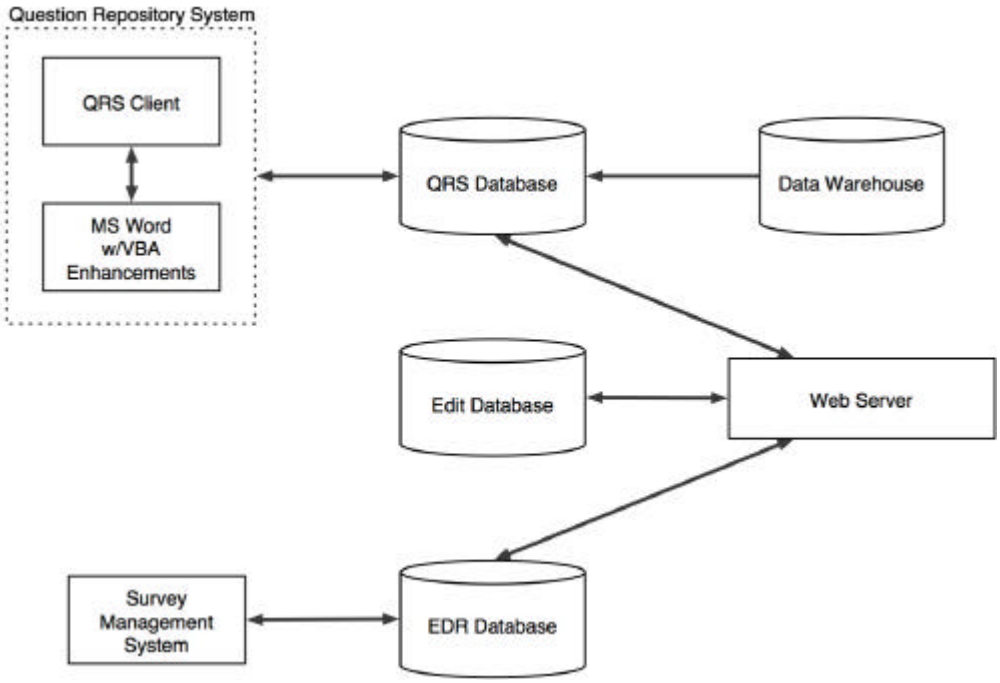

Figure 1: Simplified diagram of NASS' Electronic Data Reporting System

23. The EDR Web server runs Apache software and relies on PERL common gateway interface (CGI) scripts to query the necessary databases to dynamically construct and deliver Web questionnaires. The CGI scripts perform three key functions: (1) confirm that a respondent-provided identification number exists in the EDR database, (2) identify what surveys a particular respondent is to receive, and retrieve the appropriate survey questions from the QRS database to dynamically create the Web questionnaire, and (3) store respondent-provided survey data to the EDR database under the appropriate master variable names. In addition, the CGI scripts perform a series of formatting operations that serve to adjust the rendering of the Web questionnaires to conform to NASS business rules and Section 508 compliancy guidelines.

24. The EDR database stores the Web survey samples (name, address and other identification information) and the data the Web respondents provide. NASS' proprietary Survey Management System (SMS) provides the user interface for populating the Web samples and for retrieving Web provided survey data. The Web-provided survey data are then combined with data obtained from other modes and processed through post data collection editing and summary systems.

25. It should be noted that the NASS EDR system also contains various firewalls and other security measures. These topics were intentionally omitted in this paper because they are not only beyond the scope of the paper, but also to avoid disclosing any information that may jeopardize their effectiveness.

VIII. DATA COLLECTION EDITS:

26. Except for the Census of Agriculture that relies primarily on mail-out/mail-back data collection, the majority of NASS' data are collected via a mode that allows some degree of data collection edits (also known as *nearer source edits*). These edits may be automatically invoked by computer software, or employed by trained interviewers who collect data on paper questionnaires. Regardless of the mode, the purpose is the same – to collection higher quality data. More specifically, the data collection edits are meant to identify potential reporting errors and correct them, as necessary, with the respondent during the initial interview. Without data collection edits, imputation methods or burdensome re-contacts would be necessary. Virtually all of NASS' data collection edits are subsets of post-data collection edits that are executed with SAS and/or the Blaise interactive edit mode.

27. The majority of the data NASS collects are numerical values corresponding to various facts about agricultural operations. Typical examples include total acres operated, acres of corn planted, or number of head of cattle. Data collection edits associated with these data may be classified as (1) range edits, (2) balance edits, or (3) relationship edits. For all types of edits, upon being triggered, respondents are provided with a brief statement explaining the edit failure, along with instructions on how to correct or verify the situation.

28. Range edits check to ensure reported values fall within an acceptable (or most likely) range of answers. The acceptable range may also include requiring an answer (i.e., non-null). These edit checks are most commonly used to ensure a reported crop yield falls within a reasonable range. Since there is much variation in agriculture across the United States, the acceptable ranges often vary by state. Range edits also guard against misplaced decimals when CATI interviewers record responses, as well as misreporting units. The latter situation can occur when respondents mistakenly report a yield per acre instead of total production for a crop (or vice versa).

29. Balance edits are used to ensure that a total equals the sum of its parts. That is, respondents are sometimes asked to provide answers to individual components as well as the total of those components. The balance edit will check to see if the total does indeed equal the sum of the individual components. Alternatively, the total may be automatically generated (in CATI or on the Web) and the respondent is asked to verify the total. Generally this type of edit is reserved for situations where the highest amount of accuracy is needed.

30. As the name suggests, relationship edits check if reported data abide by either necessary or likely relationships. An example of a necessary relationship is the number of harvested acres of a certain crop cannot exceed the number of planted acres of that crop. An example of a likely relationship is if a farm has livestock, it should also report livestock expenses. Relationship edits essentially check to make sure the data "make sense." Some relationship edits may involve answers to many questions, some which had been provided much earlier in the interview. For example, the initial questions in many NASS surveys collect the total acres in the farming operation; this serves to define the reporting unit for the remainder of the questionnaire. The survey may proceed to ask for the number of planted (or harvested) acres of many specific crops, and at some time many questions into the interview, the sum of all of the reported crops may exceed the total acres operated. This would trigger an error and the respondent would be requested to correct the situation, which could literally involve a dozen questions.

31. Every data collection edit is also classified as either *hard* or *soft*. Edits classified as "hard" identify reported data that are virtually impossible to occur. One example of such an edit is if a respondent claims to harvest more acres of a particular crop than he/she planted. Data failing a hard edit must be corrected before proceeding with the interview. For the example, the respondent may either correct the reported planted acres, the reported harvested acres, or both. Soft edits identify reported data that are unusual, but still possible. Soft edits do not require changes to reported data (i.e., the error may be suppressed).

IX. ELECTRONIC DATA REPORTING EDITS

32. The EDR system relies on an edit repository for the edit checks that act on respondent-provided data. These edits are coded in PERL and may consist of virtually any logical operations to the data. The QRS database houses this edit repository and stores the actual PERL syntax along with any error messages the respondent would receive. NASS personnel knowledgeable with PERL create the edits and error messages through a Web browser interface. The edit logic uses master variables, and once created, the system automatically applies the appropriate edits to all Web questionnaires that contain all of the necessary questions. For example, if the edit logic contains the variables A and B (e.g., to check if A < B), then this edit is applied every time the questions that contain the variables A and B appear in a Web questionnaire. Hence, edits are built at the variable level instead of for specific questions or questionnaires. This makes programming edits more efficient since each specific edit need only be built once. In addition, this approach ensures more consist editing across surveys. Although the edit repository is currently used only for Web questionnaires, it may be expanded to include CATI and post-data collection edits.

33. The Web edits run server-side; hence they can only be invoked after respondent-provided data are sent to the server. At the end of each Web page of survey questions, the respondent is instructed to click a "Next" button to proceed to the next Web page of questions. Upon clicking the "Next" button, the data entered on that Web page are sent to the server and any edits that pertain to all completed questions are invoked. If any edit failures exist, a red message in slightly larger font is displayed at the top of the screen and red asterisks are placed before the question answer(s) involved with the edit failure.

34. There is still some debate as to how data collection edits should be implemented – specifically how many, how complex the edits should be, and how should error messages and visual cues to indicate which responses are involved be conveyed to respondents. NASS has conducted a limited amount of usability testing for its Web questionnaires, and is working to conduct even more. However, in the absence of much empirical data to support one approach to implementing Web data collection edits over other methods, NASS has elected to take a fairly conservative route. To this end, very few data collection edits have been added to Web instruments. Those that have been added are generally subsets of the CATI data collection edits used by interviewers and all are "soft" edits (i.e., no changes to reported data are mandatory). The one exception to this rule pertains to numerical checks that ensure that responses that are expected to be numeric are indeed numeric. This edit is used on virtually every question and ensures no type mismatches occur with data being stored to the EDR database or for post-data collection processing. To minimize numeric edit failures, the server automatically removes dollar signs, percent signs and commas from responses (i.e., it removes non-numeric characters from an otherwise numeric response).

X. CATI EDITS

35. Since they are automated and have an interviewer present, NASS' CATI data collection edits are more numerous and complex than Web data collection edits. CATI data collection edits consist of range, balance, and relationship edits. While NASS uses both hard and soft CATI data collection edits, most are soft. Most suppressed soft edits allow interviewers to enter comments to explain the situation, which they are encouraged to do. When error conditions are violated, the Blaise software displays an error message along with a list of fields involved in the error. Navigation to these involved fields is a simple matter of double-clicking on the field name. Data collection edits in CATI are a subset of those used for post-data collection. However, compared with Web and face-to-face interviews, CATI data collection edits have a unique relationship with post-data collection processing in that suppressed soft edits are ignored by post-data collection edits. This is accomplished by setting appropriate flag variables in the Blaise CATI instrument.

XI. FACE-TO-FACE EDITS

36. Face-to-face interviews are used for large operations, other operations that have requested that mode, or for past refusals. All face-to-face interviews are conducted with pen-and-paper questionnaires, as NASS does not [yet] utilize CAPI. NASS provides thorough training to its interviewers, including detailed interviewer's manuals that explain all survey questions, as well as any relationships that typically exist between questions. This training provides the basis for face-to-face data collection edits. When responses violate the typical data relationships, interviewers are instructed to ask respondents to explain the situation. Changes are made if necessary, and interviewers are instructed to leave comments on the questionnaires if no changes are made. The comments serve to validate the reported data when they inevitably fail post-data collection edits. Compared to those data collection edits in CATI or the Web, there are fewer face-to-face edits and they generally confined to relationship edits. Since there is no automated process for applying the face-to-face data collection edits, they are certainly subject to interview variability.

XII. SUMMARY

37. NASS' complex data collection program involves hundreds of census and sample survey programs. Data are collected via traditional mail, telephone (CATI), face-to-face, and the Web. These data collection modes are implemented simultaneously for any given data collection program in order to minimize non-response. Data collection edits are utilized for CATI, the Web, and face-to-face modes in order to improve data quality; these edits are generally subsets of post-data collection edits. Problems identified and corrected during data collection minimize the amount of imputation during post-data collection processing and reduce burdensome re-contacts.

38. Since it is still relatively new, NASS is still learning how to best integrate the Web into its data collection program. One significant issue that still remains is how extensive and how complex Web data collection should be. Further research and usability testing are needed to address this and other remaining Web data collection issues.

XIII. REFERENCE

Government Paperwork Elimination Act. Pub. L. 104-13. October 21, 1998. United States Title 44 Chapter 35.
