WP.3
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (i): Editing of data acquired through electronic data collection

DATA EDITING IN A COMMON INTERNET DATA COLLECTION SYSTEM

Invited Paper

Prepared by Betty Barlow, Stan Freedman, and Paula Weir, U.S. Energy Information Administration, United States[1]

I. INTRODUCTION

1. EIA's Internet data collection systems were developed independently by each program, as reflected in the different approaches taken. As part of EIA's strategic action plan, a team was appointed in 2005 to research and design a prototype for a common Internet-based survey data collection system. The team produced a design document known as the Functional Requirements Document as the first step in a common "next generation" internet data collection system to be used by developers once funding became available. In 2008 EIA launched the first survey in the new collection system with others soon to follow. This paper will focus on the data editing aspects contained in the original functional requirements, as well as the editing aspects in the authoring tool that was also developed to render survey forms in the new EIA Internet Survey Management System.

II. THE FUNCTIONAL REQUIREMENTS DOCUMENT

A. The Vision

2. It was the vision of the Energy Information Administration (EIA) Internet Data Collection (IDC) Team that EIA would develop and implement an enterprise-wide IDC system (hereafter referred to as the IDC) to take EIA forward in its data collection approach. In this vision, future data collection tools must be flexible and versatile because surveys across EIA would be operated in a multi-mode environment (e.g., telephone, mail, Internet, email, personal interview, etc.). While EIA would encourage respondents to report through the Internet, optimizing survey response would still the over-arching objective. The vision for EIA was that the IDC be viewed as an integrated collection framework, rather than a single, independent mode of collection. Therefore, the IDC Team defined the users of the IDC as not only the respondents who submit data through the Internet, but also internal EIA users who would employ the IDC as the data capture mechanism for other collection modes. The IDC would consolidate these reporting modes in one logical location. The IDC infrastructure would provide the building blocks needed for interoperability and enable information to be shared by the survey processing system and the information dissemination system. This flow of information would contribute to the comprehension and interpretation of data by the customer. The interoperability attained would provide two or more system

[1] This report is released to inform parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily of the U.S. Energy Information Administration

components the ability: to exchange information (syntactic) and to utilize the information that has been exchanged (semantic). Inherent in this IDC vision was the idea of creating one location for the survey instrument and the authoring of the instrument, including metadata elements (e.g. survey questions) and paradata elements (e.g. survey response indicators). This vision allowed for the efficient management of survey instruments across modes. It incorporates the concept of survey form components contained in repositories that could be shared across surveys, enabling consistent processes and questionnaire documentation. In addition, it was envisioned that there would be the extensive application of usability testing prior to implementation to ensure that the survey components were user-friendly and understandable. Reusing thoroughly tested components in the repository would reduce the development and testing costs for subsequent surveys.

3. This vision reflected the collective experience of the IDC Team, whose members were drawn from across EIA. Above all, the vision recognized that the mission of the agency was best achieved when there is a firm commitment from Senior Management and a collaborative partnership exists among all offices of the organization, survey and support offices.

B. Three Part Focus

4. The primary work of the IDC Team focused on:

- Development of IDC "principles" structured around three functional areas which address the point of view of the IDC users, including the respondent, the survey Program Office, and the EIA internal corporation.
- Research of other Federal Statistical agencies' approaches to Internet data collection to determine the best practices and approaches relevant to these EIA principles. Through the discussions with these agencies, the demonstrations of their IDC, along with an understanding of EIA's diversity of surveys, systems, programs, and office cultures, the IDC Team developed the third part.
- Detailed key functional requirements for the development of a successful application at EIA based on the principles developed and the research conducted. In this context, success was defined by the value of the application to respondents, as well as internal EIA users. The value would be measurable by the number of surveys implementing the EIA IDC, as well as the usage of this reporting option by respondents.

5. The Functional Requirements Document was written for EIA's next generation of Internet data collection. The functional requirements were not intended to be a prototype, but rather a framework to be used by developers to construct a design document which satisfies these functional requirements once funding became available. The focus, therefore, was on *what* is required, rather than on the specific technical methods of *how* to satisfy those requirements.

C. Three Functional Areas and the Related Principles

6. The development of IDC "principles" structured around three functional areas. These principles were generalizations which were agreed upon by the team and used as the basis for reasoning and obtaining consensus in developing the functional requirements. These principles were approached using three points of view—the respondent point of view, the survey program point of view, and the EIA corporate point of view. For each of these view points, the IDC Team agreed upon which characteristics would be important.

7. The Respondent Point of View included:

- Ease of access and use (install, updates, sign-on, navigation, system's response, down-time, and other usability issues)
- Ready and useful help and prompt support
- Ability to report large amounts of data without keying or using direct data transfer

- 24-7 access
- Reduce burden compared to other modes of reporting
- Confidentiality protection
- Ability to interrupt and complete surveys later, save/print/forward, finalize, and transmit information to the respondent's co-workers and between EIA and the respondent
- Other incentives for respondents (What's in it for me?)

8. A number of characteristics were defined for the survey program point of view. A few of these are listed below.

- Instrument design and flexibility, cognitive issues; individual survey needs differ (paper vs. web look, skip patterns, scroll vs. multiple screens, "open"/roster forms, acceptability of blanks, automatic calculations, drag and drop visual objects in survey design)
- Reduce data entry costs, mailing costs, and costs associated with data editing and follow-up with respondents.
- Reduce reporting errors and nonresponse error, prevent mode bias (increase survey quality)
- Edits (within cells, across cells, and across reports or periods)
- Ease and flexibility of incorporating or modifying edits and monitoring edit performance within IDC
- Ease of resolving edit failures within IDC (including overrides, comments, or changing data)
- Ability to measure process performance in real time and historically (reliability, robustness, and flexibility)

9. The EIA Corporate Point of View included:

- A centralized platform
- Shared architecture and infrastructure for IDC (reduction in cost of systems' development and maintenance) and user management
- One entry point
- Common look and feel
- Security (for EIA)
- Ability to support multiple surveys
- Section 508 compliant (electronic and information technology is accessible to people with disabilities)
- Life cycle records' management

10. Using these three views, detailed questions were written regarding these characteristics to determine from other statistical agencies how and if these characteristics were accounted for in their Internet data collection application.

D. Interviews and Literature Search

11. A summary of these detailed questions was provided to the agencies prior to meeting with them. Each committee member was responsible for one or more categories of questions to ensure that answers were obtained in the meeting and that each agency was asked the same ten key questions and the answers were recorded. The write ups on these interviews became the case studies in the functional requirements document. These case studies were supplemented with a literature search that was conducted on best practices for Internet data collection applications. These practices from the case studies and the literature search were then examined and compared in defining the recommendations for EIA's IDC.

E. High Level Recommendations

12. Four guiding principles focused the work of the IDC Team. Each of these was supported by detailed requirements. In particular, the IDC should be:

1) Designed so that it is easy to implement, modify, and migrate an existing survey into the IDC; development of interdependent modules is the key design element for implementation.
2) Provide tools which enable the Program Office to have ownership of their survey application.
3) Provide value added from the respondents' point of view.
4) Promote high quality data through editing capability, user notification, clear navigation, and transparency across EIA surveys.

13. For the first principle, the team believed that the IDC should have a modular design featuring reusable code. This would facilitate development of the entire system by allowing for the prioritized development of the most important functionalities while also taking into account available resources. Parts of the IDC could be implemented without completing the entire system. This modular phased approach would facilitate the migration of existing surveys into the IDC. For example, once a survey was implemented that used range check and historical edits, the existing code would only need to be modified slightly for the next survey. This would reduce development and maintenance costs over the life of the system. The next survey would simply redefine the survey specific range values and the historical data to be used to utilize the pre-existing edits for that survey. When a survey required an edit type not already in the system, the requirement would be addressed, and then made available to all surveys. Only parameters and data sets are survey specific. When a module was adopted for a particular survey, it would be implemented in a common way, and it would have a common look and feel. This would provide uniformity for EIA's respondents, reduce implementation costs, and provide standardized approaches to collecting data on the Internet. In addition, it would not interfere with the operational and methodological requirements of a survey.

14. "Ownership of the System", the second principle was considered to be the key to success of an EIA-wide IDC. One of the main lessons learned from the case studies was that unless there is buy-in from the Program Offices for whom the system is developed it is doomed to failure. Consequently the system should not dictate how individual survey instrument or processes are conducted. The Survey Manager should have control over how the survey form is designed within the IDC framework. Best practices were included in this document to familiarize Survey Managers with the information important to implementing the best Internet survey instrument possible. Survey Managers would use the tools provided by the IDC to execute their surveys, modify their survey instruments, and edit the data using the rules and error resolutions appropriate for their particular survey.

15. Value Added for Survey Respondents was the third principle. In particular, the IDC must provide additional incentives to EIA's survey respondents or it will not be used. Successes in respondent mode conversion have most often been the result of elimination of other options. While successful for some surveys, this approach was not seen to be appropriate for all EIA surveys. As a result, the implementation plan included in the Functional Requirements Document called for the creation of "incentive reports" for respondents who use the IDC. These are essentially customized reports which respondents could access in order to compare their company's historical data with historical EIA data from the same survey. While respondents could produce some of these reports themselves by taking their own historical data and comparing it to a table or graph of published EIA data, some reports such as a Market Share Report, would have great value and would not be possible otherwise. EIA could produce reports for respondents as a value added incentive for using the IDC.

16. The fourth principle was that the IDC promote high quality data and reduce survey processing costs. It had already been demonstrated at EIA that online editing reduced reporting errors and respondent call-backs. It had also been demonstrated at EIA that these reductions lead to decreased survey costs and accelerated data production processes. Respondents could be notified at the data

collection stage of their reporting errors/discrepancies, and they could correct their data prior to submission. EIA staff would therefore have more time to focus on broader survey quality issues.

F. Detailed Functional Requirements on "Front End" Editing

17. It was recognized by the IDC Team that some types of data editing are best performed as part of the survey processing system. However, when possible, editing conducted at the point of data reporting by the respondent was known to eliminate or reduce response errors at the source. This front end editing would not only enable higher quality data, but also reduce survey processing costs directed at error resolution including call backs, error resolution, re-editing etc. Data editing in this IDC application is defined as the process of identifying data that are incorrect or anomalous and resolving data that fail the edit process. The editing which is performed in the IDC is referred to as front end editing to distinguish it from editing performed later by the survey processing system. In the same spirit of error prevention, the document specifically recommended that: the IDC provide drop down boxes, and these should be invoked by the Survey Manager wherever possible and reasonable, rather than free form text boxes, as a method for entering data. For example, in the survey instrument, when a respondent must report information from a limited, known set of choices, such as geographical locations, or measurement unit, a drop down box should be provided to eliminate keying errors and validity checks by offering a choice of only the valid selections.

18. For other survey data elements, the IDC should provide the capability of identifying incorrect responses through the use of edit rules as determined by the Survey Manager. The IDC should inform the respondent that these edit failures must be resolved before submission will be accepted. These edit failures, which a respondent is required to correct, are referred to as hard edits in the IDC. For example, this type of edit would be appropriate for fields which must be numeric, integers, required, or of a certain length. The IDC should allow two types of hard edit resolutions for survey submission to be completed. The first type of resolution is correction of the field in a manner that passes the edit rule. The second type of resolution is presence of a comment provided by the respondent regarding the data elements' validity. One of these two types of resolution would be required for survey submission to be completed. The Survey Manager would be responsible for determining the data items to be edited, the edit rules, and the resolution type to be used for each hard edit in the survey. These edit criteria would be managed through the administrative module of the IDC.

19. The IDC should provide the capability to identify and flag other data elements which appear anomalous, as identified by edit rules. These responses may actually be correct but are sufficiently anomalous that the response should be questioned. These edits are referred to as soft edits because they do not require the respondent to resolve them prior to submission. The IDC should provide the capability to identify and display soft edit failures and provide for three types of resolution for soft edit failures: correction, comment, or no action. However, the actual selection of no action by the respondent would be required for survey submission to be accepted as complete. The Survey Manager would also responsible for determining the data items to be edited, the edit rules, and the resolution type to be provided for each soft edit in the survey. These edit criteria would be managed through the administrative module of the IDC. The identification and location of an edit failure may pertain to an individual survey cell, cell-to-cell, (within the same survey form), cell-to-cell historically (previous reference periods), or cell-to cell with respondents from the same survey or other surveys. The IDC must provide the capability for performing all of these types of data locations for editing. The Survey Manager would be responsible for deciding which of the data locations are needed for editing for the particular survey, as well as to provide to the IDC any data needed for the edit that is not available directly to the IDC otherwise. These exogenous data should also be managed through the administrative module of the IDC.

20. The types of edit rules which may be invoked for an edit in the IDC should include consistency checks, range checks, completeness/required fields checks, comparability of data items and/or parameter checks, and checks involving arithmetic calculations on the data items in the locations as described above. Even though more than one edit rule may apply to one data cell, the edit rules would be executed

during the first iteration of edit runs. The edit rules and their associated parameters must be easy for the Survey Manager to modify within the administrative module. In addition, failed data items which the respondent has resolved through a data change should be re-edited to verify that the changed data item passes all the edits and does not cause other data items to fail.

III. THE EIA INTERNET SURVEY MANAGEMENT SYSTEM

A. The Authoring Tool

21. As mentioned, research of other Federal Statistical agencies' approaches to Internet data collection to determine the best practices and approaches relevant to these EIA principles had been conducted prior to determining the functional requirements. One of the approaches examined in detail was the National Agricultural Statistics Service's system for building surveys called the Question Repository System (QRS). Based on Microsoft Word, QRS is a system which allows non-technical personnel to choose the questions that go on a form from a repository of questions and produce the form with the press of a button. The focus of the QRS was to produce paper questionnaires which have a common look and feel though it also produces the web form and a form used for CATI. EIA's interest relative to a web form was centred on five aspects: 1) a lay person can generate the survey form using a relatively simple interface because all the work of producing the form was done upfront, 2) all forms have a common look and feel, 3) the system is multi-modal, 4) the system is based on the concept of code reuse since a single tool generates all surveys, and 5) there is a partnership between the IT group and the survey groups. The recognition of a need for an authoring tool became core to EIA's common internet data collection in order to enable survey managers to migrate their surveys more easily to Internet collection and maintain the Internet survey thereafter.

22. The authoring tool EIA developed is a part of the overall Internet Survey Management System (ISMS). The ISMS was designed to streamline and standardize the authoring of survey forms, the fielding of these forms over the Internet, the collection and validation of respondent data, and the processing of these data or the transfer of these data to legacy processing systems. The system allows a survey manager to build a survey form element by element, defining an element by selecting the appropriate data type, selecting a web control, and providing help text, a default value, and other properties. For selection-type elements (e.g., drop-down boxes or radio buttons), the survey manager may choose from a list of options. For grid-type elements, the survey manager may specify a set of columns, each with its own data type and web control. For each element, the survey manager may also specify edit rules that will be used to verify the respondent's submitted data. The survey manager also specifies the messages to be displayed to the respondent when a value fails the edit. The survey manager defines the workflow for elements of the instrument which governs how the respondent progresses through the survey. Examples of such workflow are the switching off or on of an element based on the respondent's response to a prior question; the linking of a set of drop-down boxes, such that the value selected in the first control governs the values available in the second control; or the automatic forwarding of the respondent from one portion of the survey to another, again based on the response to a prior question or set of questions.

23. Navigation through the authoring tool is fairly intuitive, and each page follows the same general layout making use of back, next, and reset links. Entries are automatically saved each time "next" is clicked at the bottom of the page. In addition to the navigation links, the left side of each screen displays the entire list of the steps for creating a survey with the current step highlighted in yellow. Authors can click on any step to navigate through the creation of a survey instrument. The first step requires the author to provide basic information about the survey such as the periodicity of the survey or the name of the survey through text boxes and drop-down menus. In Step 2, the author defines other instrument metadata such as the number of respondents or the publications that use the survey data, and provides the value for each piece of metadata. Three pieces of metadata are required: the confidentiality of the data, whether reporting is mandatory or voluntary, and the estimated burden hours to a respondent. Additional pieces of metadata are the author's option. The metadata that are entered in this step are compiled into a table that the author can view at the bottom of their screen.

24. In Step 3, the author describes the structure of the survey instrument defining it as either continuous or hierarchical. A continuous structure presents the questions in the order specified without levels, while a hierarchical structure presents the survey in levels, allowing the author to create schedules, sections, parts, and any other level to the survey. The default or standard EIA structure consists of schedules, sections, parts, and questions.

25. In Step 4 the author creates the detailed elements of the instrument. Step 4 not only contains an initial screen to enter general element information, but the step also contains four sub-parts: a) create or change an option list; b), create or change multiple record control (for tables or grids in the survey instrument); c) create or change element specific edit rules; and d) create or change element related metadata; and e) create a fixed or variable grid, or a table. Authors can create new elements or copy and change other elements from the survey or from another survey. Authors can also choose to have additional text show up in printed surveys. For example, some surveys make use of certain codes that are found in the appendix of printed instructions but appear as drop down menus in the electronic survey. The printed version of the survey would refer reporters to the appendices, but this reference is not necessary in the electronic survey. The author can also make use of the text area to enter basic or brief help text for the element or extended help text for lengthier explanations.

26. If response to certain elements is required, or is required to be within certain parameters, the author defines in step 4c the rules for errors and warnings as shown in Figure 1. An **error** (hard edit) will prevent the respondent from proceeding to the next element of the form while a **warning** (soft edit**)** will display a message but allow the respondent to continue filling out the form. The author can also choose to have messages that are only displayed to inside users such as the survey manager but not to the respondents. The edit rules entered by the author in this step are compiled into the table shown at the bottom of the page of the authoring tool. To modify an existing entry, the author finds the desired element in the table and updates the information. Authors can not, however delete, rules but are provided the ability to turn a rule off if desired.

27. While Step 4c creates element-specific edits, Step 5 creates inter-element rules that operate based on the response for another element. The author chooses the desired element from the drop down box to populate automatically with the primary elements created in the previous step. The author has the ability to choose an element from a different survey instrument with a temporal qualifier to use as comparison data in the edit rule.

28. The IDC Implementation team recommended that seven edit types be pre-defined in the authoring tool. Each edit type is associated with preconditions (in some cases required and in some cases optional), the condition, and an optional tolerance around the condition. This approach was motivated by the U.S. Census Bureau's StEPS edit module. The edit types were based on the most common current EIA edits which included: 1) positive value (if not null); 2) required item; 3) one- or two-sided range on element or ratio of two elements; 4) balance of elements; 5) prior-no-current/current-no-prior; 6) list directed; 7) free form for more complex or custom rules. It was not expected that all edit types would be available in the initial operational capability of the system but the system would be expanded from initial capability to include all of the most common edit types.

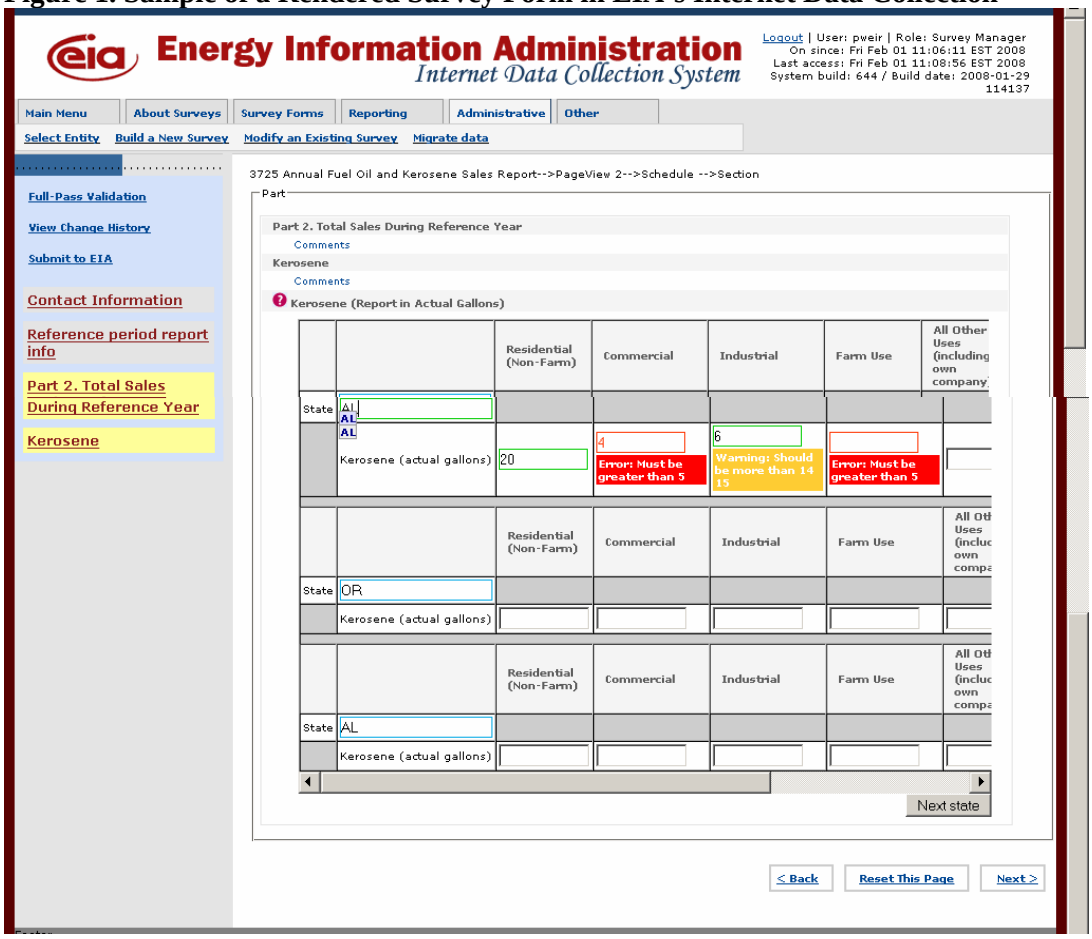
29. In Step 6 the author defines the element related workflow and in Step 7 the author assembles and tests the rendered survey form. The system automatically routes the author to an entry list screen for their survey instrument.

B. The Rendered Survey Form and Reporting

30. The rendered survey form maintains the same general look and feel as in the authoring tool. The agency branding is at the top of the page along with author survey-level metadata defined in the authoring tool. The back, next and reset links are at the bottom of the page, and the navigation pane on

the left provides the sequential reporting steps. The step that the respondent is currently on is highlighted in yellow. In the illustration shown in figure 1, the author had designed the form using a variable grid to allow the respondents to enter data for related element more easily. In this made up example, the author had chosen to have the edits invoked as the respondent tabs to the next cell. The author had also selected the web control option for the State element as autocomplete. As the respondent enters the first letter of the State, the possible State abbreviations that begin with the same first letter are provided. The respondent selects from the list the appropriate one. Two severity levels of edits were implemented in this example, errors and warnings. The variable grid design selected by the author creates a new row at the bottom of the grid when the respondent selects "next state" at the bottom.

Figure 1. Sample of a Rendered Survey Form in EIA's Internet Data Collection



31. Once the respondent has completed the form, the respondent selects the full pass validation link in the navigation pane. This step must be performed before submitting the form to EIA.

IV. CURRENT ISMS STATUS

32. The first survey to be implemented in ISMS is scheduled for initial operational capability at the end of the first quarter of 2008. The first survey has unusual characteristics such as the length of the survey and the option for respondents to attach other material or supporting documents to the response. Two more surveys are scheduled for implementation soon thereafter. It is expected that the authoring tool will be enhanced during this process to include the selection of standard text such as the legislative authority for conducting the survey, pre-defined drop down boxes such as state codes, or product names. Initial operational capability for edits is fairly limited in that the only choices are: required, greater than, greater than or equal to, less than, less than or equal to, or equal to a constant value provided by the

author. Compound rules and the inter element edit capability in step 5 are not yet operational. While not yet available, it is expected that the respondent's historical responses can be used as values for the edits. In the short run, though, migrating surveys requiring more sophisticated edits than available in the authoring tool are being directly provided outside of the authoring tool.

V. SUMMARY AND CONCLUSIONS

33. The functional requirements document took roughly two years to develop. It documents the common vision for developing a corporate internet data collection system. By focusing on overarching objectives and internet survey best practices, consensus was reached on what the corporate requirements were, paving the way to a successful corporate change. This document has become a reference guide for both the system's developers and the survey authors for determining what features to implement. Other suggestions are provided to survey authors such as making sure the edit failure messages are clear as to the action to take from the respondent's viewpoint.

34. The development of the ISMS system has progressed far in one year. The inclusion of an authoring tool as a part of the system has addressed the four guiding principles recommended for a corporate system. In particular, it enables the program offices to have ownership of their survey applications, allows for the program offices to migrate surveys into the ISMS, and promotes high quality data through the editing capability, clear navigation and user notifications. The third principle of valued added from the respondent's view is also being addressed through features such as the import function and immediate edits (to reduce follow-up phone calls later). The historical data provided to the edits could further be used to create special reports for the respondent to provide further incentive for reporting through the ISMS and encourage change from the respondents in how they think of government reporting.

References

Anderson, Amy E.; Cohen, Stephen; Murphy, Elizabeth; Nichols, Elizabeth; Sigman, Richard; Willimack, Diane K.; (March 2003) "Changes to Editing Strategies when Establishment Survey Data Collection Moves to the Web," Paper presented to the Federal Economic Statistics Advisory Committee, Bureau of Labor Statistics, Washington, D.C., WEB LINK: http://www.bls.gov/bls/fesacp2032103.pdf

Arbues, Ignacio, Gonzales, Manuel, Gonzalez, Margarita, Quesada, Jose, Revilla, Pedro. (2005). "EDR Impacts on Editing," UNECE/Eurostat, Work Session on Electronic Data Reporting, WP No. 27, May 2005, Weblink: http://www.unece.org/stats/documents/2005.05.sde.htm

Beckler, Daniel. (2004). "The NASS Question Repository System," In Proceedings of the American Statistical Association, Section on Survey Methods Research. Alexandria, VA: American Statistical Association.

"Internet Data Collection Requirements Document" EIA internal document prepared for the EIA Strategic Plan Goal 4 Subcommittee, October 2006

"ISMS User Guide for the Energy Information Administration Survey Authoring Tool", EIA internal document, July 2007.