INTERNET DATA COLLECTION AT THE U. S. CENSUS BUREAU

Howard Kanarek and Barbara Sedivi[1]
U. S. Census Bureau

Introduction and Background

The most common methods of data collection for surveys and censuses are postal mail, personal interviewing, and telephone interviewing. The growing number of respondents with access to the Internet introduces a new data collection alternative that is likely to become increasingly important in the future. Like computer assisted telephone and personal interviewing, computer assisted self-interviewing using the Internet permits an interactive exchange with the respondent through intelligence built into the computer application. While promising, Internet surveys also face a variety of challenges in survey coverage, in survey design, in security of confidential information, and in mastery of new and rapidly changing technologies.

At the U. S. Census Bureau, development and implementation of new collection technologies has been led by its Computer Assisted Survey Research Office (CASRO) working in collaboration with other research and production divisions. The initial focus of this office was implementation of computer assisted telephone and personal interviewing (CATI and CAPI) for demographic surveys. Concurrent but smaller efforts were initiated on other new collection technologies, including touchtone data entry, voice recognition, and optical character recognition of facsimile returns. Among these secondary goals was development of computerized self-administered questionnaires (CSAQ) for economic surveys and censuses. The first approach to CSAQ at the Census Bureau was to offer intelligent questionnaires on diskettes, resembling the Department of Energy PEDRO system and commercial software used to prepare IRS tax forms. With the growth of the Internet, the CSAQ program was expanded in 1996 to encompass electronic questionnaires transmitted over the Internet, especially those employing the World Wide Web.

This paper summarizes the requirements, procedures, and technical options chosen by the Census Bureau to develop its Internet CSAQ capabilities. These choices were based on the CSAQ pilot tests listed in Table 1, which began in 1993 (Harley et al., 1998 and Ramos et al., 1998). Note that the early tests employed the disks-by-mail technique while the later ones employed Web surveys or concurrent disk and Web options. The sample sizes given in Table 1 are the initial test samples. Production use of CSAQ methods followed the tests with larger samples. For example, the number of companies utilizing CSAQ to report in the Company Organization Survey has grown from 89 in 1993 to 1,272 companies in 1998. These 1,272 companies provide data for 283,526 establishments, which represent 69.4 percent of the establishments in this survey.

[1] This paper reports the general results of research undertaken by the U. S. Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the U. S. Census Bureau.

Table 1: CSAQ Tests

SURVEY	MODE	TEST CASES	RESPONSE RATE	PLATFORM (OS)	SOFTWARE	START DATE
1993 Survey of Surveys	Diskettes	52	100%	DOS	CASES	Feb. 1994
1993 Company Organization Survey	Diskettes	89	77%	DOS	Clipper and C	Mar. 1994
1994 Survey of Industrial Research and Development	Diskettes	100	53%	DOS	CASES and EIA PEDRO System	Feb. 1995
1994 Annual Survey of Manufactures	Diskettes	15	88%	DOS	Clipper and C	Jan. 1995
1996 Survey of Industrial Research and Development	Web	50	68%	Windows 95/ Web Browser	HTML/ JavaScript	Apr. 1997
1996 Survey of Industrial Research and Development	Diskettes	2,552	16% (not pre-screened)	Windows	Delphi	Apr. 1997
1997 Economic Census - Retail	Diskettes	550	95%	Windows	Delphi	Dec. 1997
1998 Library Media Center Field Test	Web	924	1.0%	Windows 95/ Web Browser	HTML/ JavaScript	Oct. 1998
1998 Company Organization Survey	Diskettes and Web	48	27%	Windows 95/ Web Browser	Delphi Executable	Jan. 1999

CSAQ Functional Requirements

A working group representing many areas across the Bureau established CSAQ Functional requirements for the Census Bureau. The key requirements are summarized below. Many of these requirements, especially those near the top of the list, are similar to those previously established at the Census Bureau for CATI and CAPI.

- Reproduces the survey questions and instructions in an electronic format
- Allows response entry through data keying or data entry controls (radio buttons, check boxes, pick lists, etc.)
- Provides both sequential or non-sequential navigation
- Incorporates auto-filled and auto-calculated fields
- Performs branching and skipping of items based on response values
- Uses edit tests to verify the response values for accuracy, integrity, completeness and shows error or warning messages
- Uses field checks to test the validity of the response type (alphabetic, numeric, date, phone

number, etc.) and shows error or warning messages

- Provides final survey review facility to check for missing responses and unresolved edit failures
- Allows respondents to print their response values
- Offers a means for respondents to provide free-form comments
- Provides context-sensitive help and general information about the survey
- Allows the respondent to save the data already entered and to retrieve these data upon resuming the reporting session at some later time
- Produces output that can be integrated with the survey processing system
- Passes Usability testing
- Compatible with Web browsers Netscape Communicator 4.0 and higher or Microsoft Internet Explorer 4.0 and higher
- Requires zero (or minimal) user configuration

Additional functional requirements were found necessary for Census Bureau surveys of business establishments conducted by the Bureau's Economic Directorate. These were:

- Accommodates large amounts of pre-loaded data
- Imports and exports respondent's data files
- Permits authoring of the survey once for both diskette and Web applications
- Permits, interruptible, survey completion in sessions

These last requirements were not necessary for the one institutional CSAQ survey thus far undertaken by the Census Bureau, the Library Media Center Survey.

Web CSAQ Design and Usability Testing

Based on these emerging requirements, we began development of CSAQ surveys. The initial design in the early 1990's used DOS diskettes and largely emulated the DOS CASES instruments used by the Census Bureau for CATI. When CSAQ development switched to Windows in 1996, we employed Graphical User Interface (GUI) features using Windows development tools. As much as possible, the same GUI features were used in the Web CSAQ design. Additionally, each Web CSAQ was created through a series of iterative trial and error steps followed by usability testing and/or expert review (Murphy et al., 1999; Nichols et al., 1998; Sweet et al., 1997; and Zukerberg et al., 1999). The following sections summarize the results of this effort in four design areas: access and submission; navigation; edits; and help information. The results presented here are selected from a much larger set of guidelines and recommendations (and individually supporting references) available from the authors. Topics for which recommendations are available that are not included here are use of pre-loaded historic data; importing data from respondent's files; and single authoring of forms for multi-modal (paper, diskette, Web, CATI) applications.

Access and Submission

There are several ways a Web survey can be presented to respondents.

1. The respondent accesses the CSAQ on the survey agency's Web site, completes it interactively in one session with their Web browser, and submits the data once at the end. This design works best with simple, short questionnaires.

2. The respondent accesses the CSAQ on the survey agency's Web site and answers its questions interactively with their Web browser, but they may take several sessions to complete the form. When a respondent needs to exit the questionnaire, he or she clicks a "Quit Temporarily" button and the data are saved on the survey agency's server. The next time the respondent accesses the survey, the saved data are loaded back into the CSAQ application. This design is appropriate for long questionnaires that respondents often cannot complete in one session.

3. The respondent accesses the CSAQ on the survey agency's Web site but instead of answering its questions with their Web browser, they down load an executable CSAQ, often with pre-loaded data, to their hard drive. The respondent completes the CSAQ off-line, perhaps over an extended time or multiple questionnaire waves. When answering is complete, they return to the survey agency's Web site and upload their reported data. This design is best when the survey employs extensive pre-loaded (historical) data or asks the respondent to import data from their own files.

Navigation

There are basically two ways to navigate through a Web CSAQ. One is called scrollable, where the respondent moves down one long page containing all of the survey questions using the scroll bar to the right of the screen. The second is called screen-based, where the form is broken into separate pages accessed by Next and Previous (or Back) buttons at the bottom of each page. Plain HTML only permits scrollable navigation.

Usability testing suggested the following navigation guidelines.

1. Be consistent in the type of navigation used within a particular questionnaire. Avoid mixing scrollable and screen-based navigation in the same questionnaire. This will help to keep the respondent from getting lost and from missing items.

2. In the scrollable approach, have the respondent navigate between fields by point and click instead of by using the tab key. This way the respondent will have more control over the position of the cursor. It will help prevent respondents from entering data in the wrong field. It also will avoid confusion when the tab key moves the cursor to hyperlinks and icons.

3. Screen based design is preferable if skip or branching patterns are required for the questionnaire, or if controlling the respondents flow through the questionnaire is important.

For screen based navigation, have the respondent navigate through the pages by using Next and Previous (Back) buttons at the bottom of the screen and/or by use of a continuing menu bar for the entire questionnaire at the left of the screen.

4. When a menu bar is used, it should list all the questions in the questionnaire, not just the questionnaire sections, scrolling or expanding if necessary. The menu bar should highlight the item corresponding to the respondent's current location and indicate which questions have been completed.

Edits

Edits may be written into a Web CSAQ questionnaire to inform the respondent that an entry may be incorrect. Usually this is done with a pop-up window describing the problem or error. The pop-up window should be displayed close to the field in error. This is easier to do in a page-based approach than in a scrollable approach. In a scrollable form, it may be necessary to indicate the item number with the error message to let the respondent know where the error is.

Edits should occur automatically when the respondent makes an entry or presses the Next button or reaches a fixed point in the questionnaire. Allowing the respondent to decide when to run the edits by clicking an Edit button takes the risk that the edits may never be run.

Edits may generate two types of messages. One is a warning, which the respondent may ignore if he or she chooses. The other is an error message which requires an action by the respondent. Often the respondent cannot move to the next item or submit the completed data until the error is resolved.

Edits may be active or passive. An active error message window may include an OK button and a Cancel button. If the respondent clicks OK, the erroneous entry is cleared and left unanswered. If the respondent clicks Cancel, the error message window closes but the original entry remains unchanged. Since respondents may click OK or Cancel simply to remove an annoying pop-up window, they may be unaware of the consequences of their choice. A passive error message offers only an OK button which closes the error message window but leaves the erroneous entry untouched and requiring a change. We discovered through usability testing that the passive error message is preferable.

Editing of the Web CSAQ data varies depending on the software used to edit the form and where it is run.

1. If a plain HTML form without JavaScript is chosen, no edits can be included in the form. In this situation, edits can be run at the survey agency's Web server when data is submitted. This software option is nevertheless chosen by some to ease download time or to make the form more universally accessible.

2. Edits can be incorporated within CSAQ questionnaires near the applicable data fid if they are written in HTML with JavaScript or Java applets or if they are written as executable

programs. This is true both for simple field validations, such as range checks or alpha/numeric checks on individual entries, and for more complex edits, such as those for missing entries, inter-item inconsistencies, or checks against historical data.

3. The same edits may be placed both within the form and at the end of the form. This is recommended practice so that the respondent will have a second chance to resolve errors not fixed previously. For edits placed at the end of the form, provide an easy way for the respondent to return to the field that is in error to correct it.

Help Information

The general model of help information in a Web application assumes that when the user (survey respondent) encounters difficulties, he or she clicks on a help icon or chooses a help menu item and the system opens a window providing the needed information or a means to find it. As with edits, pop-up windows for help information are not possible if plain HTML is used.

Perhaps the first consideration in designing help information for a Web survey is to decide whether help buttons or help windows are really the best means of providing the needed information. In usability tests at the Census Bureau we found respondents rarely used help buttons, help links, or help icons. Thus, if it is important for a respondent to see specific information, do not place it in a pop-up window. Instead, put it directly in the form, either at the beginning or where that information is needed. Help windows are more appropriate for optional information, information of lesser importance, or to clarify matters for respondents who are having difficulties. The following are some examples of these guidelines.

1. General information about the survey, such as its purpose, burden hours, legal status, level of confidentiality, and person to contact for more information are important to have available for the respondent but not essential for answering. This information can be made available to interested respondents through an icon at the start of the form that produces a pop-up window for respondents who want information on these matters.

2. Information about how the CSAQ form works is essential for answering and may be made available before the first question in a concise bulletinized list to be seen by all. This list may include such matters as general navigation instructions and explanations of help information icons and error messages. Usage information for specific actions taking place later in the form, such as how to perform edits at the end of the form, should be provided at the point where the respondent will first use them. As another example, instructions explaining how to respond to a radio button field should be placed in italics immediately before the first use of radio buttons.

3. Item specific information can be made available either through specific instructions incorporated in that item when crucial or through a help icon beside the related item that triggers a pop-up window if less critical. Information to include in the pop-up window when needed are the purpose of the item, definition of the terms used, and what to include or exclude in the response to that item.

The Architecture of Internet Data Collection Systems

The application of computer methods to survey data collection began with computer assisted telephone interviewing (CATI). As the Census Bureau developed additional methods of data collection such as computer assisted personal interviewing (CAPI), touchtone data entry, and diskette CSAQ, the general architecture of computerized survey information collection was expanded. This section focuses on three major areas of the architecture that were addressed for Internet data collection. They are the questionnaire, case management, and security.

Questionnaire

The functional requirements for Internet questionnaires discussed in earlier sections describe an interactive application where interview questions are presented to the respondent and actions are taken based on the responses. For example, an edit may compare a response to previously reported data, and ask the respondent to review the response for accuracy. The questionnaire may perform calculations as values are entered in a table or present help information upon request. The model most commonly used on the Internet to program interactive applications is to send the user's entered data to the agency's Web server and return a Web page to the user based on the data. This server access causes a delay and its duration is dependent on several factors including the performance of the Internet, Internet service provider, and Web server. A questionnaire designed using this model imposes annoying interruptions in completion of the form whenever an interaction with the server is required. This can be frequent for questionnaires. Usability tests indicate that checking each response as entered is preferable to having the respondent enter large numbers of responses before they are checked as a group. For this reason, the Census Bureau pursued an approach of using intelligent Web pages which execute on the respondent's computer for its questionnaires to minimize Web server accesses.

The Internet consists of heterogeneous client hardware and software. The software or browser supports published and de facto standards which allow Web pages to be displayed and execute on the client computer. Microsoft's Internet Explorer and Netscape's Communicator are the dominant browsers having the majority of market share. Both browsers support World Wide Web Consortium HTML recommendations, which provide a high level of compatibility for the display of textual and graphical information. Unfortunately, a similar standard does not exist for programming Web pages. The alternatives are JavaScript, VBScript, and Java. JavaScript is a scripting language that can be embedded in HTML. Explorer and Communicator support JavaScript, but their implementations are not fully compatible and applications that use JavaScript need to account for the differences. The usual techniques to avoid incompatibilities are to develop instruments for each browser type and version or to program conditional statements for each browser. VBScript, based on Microsoft's Visual Basic, is also a scripting language, but is only supported by Explorer. Java is an object oriented programming language used to create programs or applets that run on browsers that support the Java platform. It offers the full functionality of a programming language. Explorer and Communicator both provide the Java platform, but their interpretation of Java applets is not always consistent. Alternative methods for distributing questionnaires on the Internet is to have respondents download an executable, plug-in, or helper application using their browser. The software is installed and executes on the respondent's computer and the survey data is submitted via the Internet to the server. However, the download and installation of software may impose a burden on the respondent.

For its initial research efforts, the Census Bureau used questionnaires programmed with HTML and JavaScript, and a downloadable executable originally developed for CSAQ diskettes. At the time, Java was considered a developing technology. It did not offer the required stability, and was supported by only the newest browser versions limiting coverage. Based on survey requirements and usability testing, the Census Bureau also decided against a model where each response is sent to the server and Web pages are dynamically generated based on the response. The questionnaires programmed in HTML and JavaScript worked well, but development was tedious because JavaScript does not provide the features of a true programming language. To ensure compatibility, testing was performed across browsers. We feel that Java now offers the most potential for developing Internet questionnaires in the future. Since Java is an object oriented programming language, a set of reusable objects describing questionnaire functions can be designed. Once developed, these objects can be used to facilitate the programming of additional survey questionnaires. Ideally, an off-the-shelf GUI product will become available in the future to simplify the programming work.

Case Management

The Census Bureau like other survey organizations has considerable experience developing case management systems for CASIC technologies. Case management systems are used to distribute cases for data collection, maintain status information during the data collection process, store respondent data as it is collected, and forward the data for processing. Examples at the Census Bureau are the master control system that provides overall case management for all technologies, and the CAPI case management system which performs these functions at the technology level. The functional requirements generated for Internet data collection case management showed great similarity with previously developed case management systems. These systems use a relational database, specifically Oracle. Since Oracle met all of the functional requirements for the Internet, it was also selected as the technology solution for Internet case management.

A major requirement of the Internet case management system is to deliver and receive information from respondents. To accomplish this, the database maintains information for each respondent. This includes descriptive information for each respondent such as a unique identifier and the respondent’s name. The database also must store information that is used to initialize the Internet questionnaire and the data entered by the respondent. Depending on the survey, the form is sometimes filled with the respondent’s address and prior period information. The survey may also allow the respondent to quit temporarily by storing information on the server. This information is stored in binary large object (BLOB) fields in the database, and is merged with the Internet questionnaire by a Java servlet on the Web server. An additional requirement for Internet case management is to track respondent accesses to the survey. This information is stored in the database and includes events such as initial survey access, temporary save, and final submission. Information is also logged about the respondent’s operating system and browser type and version.

Security

Since the Internet is a public network, security vulnerabilities exist. They include the following: (a) eavesdropping, i. e., intermediaries can listen in on private conversations; (b) theft, data stolen during the course of transmission or from a computer or network; and (c) impersonation, a sender or receiver using a false identity for communication. The Census Bureau needed to address these issues to provide respondents with a secure and private method to use the Internet for survey data collection.

For the majority of Census surveys data confidentiality is protected by Title 13 and Title 15 of the US Code. The public's perception of the confidentiality of this data also influences the response rate for Census Bureau surveys and the Bureau's ability to collect accurate information from households and businesses. Since the security vulnerabilities of the Internet have been widely publicized, we devoted considerable time and effort to development of a security solution for Internet data collection that meets Census Bureau and Federal standards.

Under the Information Technology Management Reform Act (Public Law 104-106), the Secretary of Commerce approves standards and guidelines that are developed by the National Institute of Standards and Technology (NIST) for Federal computer systems. These standards and guidelines are issued by NIST as Federal Information Processing Standards (FIPS) for use government-wide. NIST develops FIPS when there are compelling Federal government requirements such as for security and interoperability and there are no acceptable industry standards or solutions. Aside from security standards the Census Bureau requires for internal applications, it also must comply with the FIPS PUB 46-2 Data Encryption Standard and the FIPS PUB 140-1 Security Requirements For Cryptographic Modules.

Security for Internet data collection had to be addressed at three levels: (1) the security of communication between the respondent and the Census Bureau; (2) the security of respondent data at the Census Bureau, and (3) the security of the Census Bureau network. Strong encryption provides for the security of data on the Internet. A product from VeriSign, an Internet security company, called the VeriSign Global Secure Site ID provides 128 bit RC-2 or RC-4 encryption and assures respondents of the authenticity of the Census Bureau Web site. The VeriSign product works with Microsoft's Internet Explorer and Netscape's Communicator version 4.0 or higher. To protect data at the Census Bureau, the respondent is required to authenticate to the data collection Web server. This server runs Netscape Enterprise Server software. Authentication is accomplished using Netscape's Directory Server software, which supports the lightweight directory access protocol, a standard for directory services. In addition respondent data is stored on a database server that resides on a trusted network. The Web server and database server are Sun Solaris systems. Finally, the Internet data collection Web server and Census Bureau networks are protected by a firewall. Collectively, these technologies provide a secure method for the Census Bureau to collect data on the Internet.

Conclusion

This paper has summarized what we at the Census Bureau have learned from research and testing Internet CSAQ applications to date. Since Web data collection is in its infancy, this is only the beginning. As Web technology matures, programming languages based on Java or XML may be developed for Web survey design, most browsers will handle code in a similar manner, and security may be enhanced by an infrastructure of digital certificates and perhaps iris scanning. Guidelines for Web questionnaire design will be further tested, standardized, and documented. With these advances and increasing Web skills in the general public, respondents will find Web questionnaires increasingly easy to use. The ease of use and intuitiveness of a Web CSAQ is important since we do not have the luxury of training CSAQ users. The Web also offers the opportunity to use graphics, audio, and video to improve the overall interview experience for the respondent.

References

Harley, D., Sedivi, B., Nichols, E., Kanarek, H. and Rogers, R. (1998), *"Internet CSAQ System Goals and Requirements,"* U. S. Department of Commerce, U. S. Bureau of the Census, Economic Planning and Coordination Division.

Murphy, B., Marquis, K., and Roske-Hofstrand, R. (1999), *"Census 2000 Internet Questionnaire Usability Testing: Initial Findings and Recommendations (Phase 1),"* U. S. Department of Commerce, U. S. Bureau of the Census, Statistical Research Division.

Nichols, E., Tedesco, H., and King, R. (1998), *"Results from Usability Testing of Possible Electronic Questionnaires for the 1998 Library Media Center Public School Questionnaire Field Test,"* Human-Computer Interaction Memorandum Series #20. U. S. Department of Commerce, U. S. Bureau of the Census, Statistical Research Division.

Ramos, M., Sedivi, B., and Sweet, E. M. (1998). Computerized self-administered questionnaires. In: M.P. Couper, R.P. Baker, J. Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nicholls II, & J.M. O'Reilly (eds). *Computer assisted survey information collection* (pp389-408). New York: Wiley.

Sweet, E., Marquis, K., Sedivi, B., and Nash, F. (1997), *"Results of Expert Review of Two Internet R&D Questionnaires,"* Human-Computer Interaction Report Series #1, U. S. Department of Commerce, U. S. Bureau of the Census, Statistical Research Division.

Zukerberg, A., Nichols, E., and Tedesco, H. (1999), *"Designing Surveys for the Next Millennium: Internet Questionnaire Design Issues,"* U. S. Department of Commerce, U. S. Bureau of the Census presented at 1999 AAPOR conference in Clearwater, FL.