A DISCUSSION OF DATA COLLECTION VIA THE INTERNET

Elizabeth Sweet and Chad Russell, US Bureau of the Census
Elizabeth Sweet, Bureau of the Census, 4401 Suitland Road, Washington, DC 20233
The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Key Words: Computerized questionnaires

1. Introduction

In recent years there has been a trend toward greater automation in the survey data collection process. Paper questionnaires used in personal and telephone interviews have been replaced in many instances by CAPI and CATI instruments. Automating the questionnaire has yielded better data quality, reduced respondent burden, and quicker survey timing.

The increasing public awareness and use of the Internet, in particular the World Wide Web, presents new opportunities for persons designing computerized questionnaires. Issues such as security, respondent coverage, and perception will determine how the Census Bureau will take advantage of this technology. This paper will examine these issues, and discuss some of the Internet data collection methods currently being reviewed.

2. What is the Internet?

The Internet loosely refers to a global network of computers that communicate with each other via the Transmission Control Protocol/Internet Protocol (TCP/IP) suite of protocols. A user at one computer, or "node," can access services provided by other nodes on the network (client/server computing). Such services include remote login (TELNET), file transfer (FTP), electronic mail, and hypertext (WWW). Using these various services, individuals and organizations can set up "Internet sites" - virtual places where a wide variety of information can be made available to anyone on the network.

The Internet is constantly growing and changing, as more and more individuals and businesses rush to establish an on line presence. It is estimated that there are currently about 13.5 million users worldwide, and that the number of Internet "nodes" is doubling in size each year (Quartermann, 1994). This growth has been sustained since 1988, and is expected to continue since even in the most densely networked countries, there are only about 8 "connected" computers per 1000 people.

Much of this tremendous growth is due to the popularization of the World Wide Web (WWW), which enables users to browse text, images, and audio and video files through a simple graphical interface. Indeed, when many people today use the term Internet, they are referring to the WWW.

Information travels on the Internet in both directions. Not only do survey organizations like the Census Bureau use the Internet to publish survey results, but they can also use it to collect information from the populations they survey. The feasibility of data collection via the Internet depends largely on whether the target populations are "connected," and if so, whether they are willing to use the Internet to report data.

3. Who is connected to the Internet?

Estimates of Internet access vary. A 1994 study by Times Mirror showed that only approximately 31% of U.S. households owned a computer, and although the Internet is growing, the number of computers connected to the Internet, whether that be through an on-line provider, direct access or modem is limited. In addition, computer owners are not representative of the U.S. population. "College graduates were 2.5 times as likely to have a home PC than were high school graduates; households with an income of more than $50,000 were five times as often equipped with a PC than were those earning below $20,000." (Oppermann, 1995) In contrast about 98% of large businesses and 63% of small business (with 100 or fewer employees) have PCs (Ogden Government Services, 1993). A 1994 survey by Forrester Research showed that two out of three large companies in the U.S. already had some kind of Internet connection. (Groenfeldt, 1995) Thus the use of on-line and computerized questionnaires might be more appropriate for business or establishment research, especially when the target population includes large businesses.

4. Preference for Internet Survey Reporting

In October 1994, the Bureau of the Census (BOC) mailed a screening questionnaire (Survey of Potential Computerized Self-Administered Questionnaire Respondents) to large companies with one million dollars or more in expenditures in the BOC's 1993 Industrial Research and Development (R&D) Survey sample. Results from this questionnaire were used when selecting a sample of companies to receive a diskette-based Computerized Self-Administered Questionnaire (CSAQ). Approximately 500 screener questionnaires were mailed. Of the 177 mail-return screener respondents that were willing to use a diskette-based CSAQ:

-56% (100) had a modem

-69% (123) reported having some type of communications capability (communications software, on-line provider access and/or Internet access)

-29% (52) indicated no communications capability

-1% (2) left all communication capability questions blank

Of those 123 R&D screener respondents who had electronic communications capability, 37% (46) had either direct access to the Internet or e-mail access. This 37% figure is smaller than the Forrester Research estimate. Nevertheless, of those 46 companies, approximately 72% were willing to receive a CSAQ electronically and 83% were willing to electronically return a completed CSAQ. (Sweet and Ramos, 1995)

More recently, a voluntary survey was included on the back of the transmittal letter for the Bureau of the Census 1995 Company Organization Survey (COS). The purpose of this survey was to gather electronic reporting preferences and capability for CSAQ, Electronic Data Interchange (EDI), and Internet. Approximately 20,000 questionnaires were distributed to primarily large companies. Because of the voluntary nature and perhaps due to the placement on the back of a transmittal letter, only 1295 companies responded as of April 23, 1996. Approximately 12.9% of those companies claimed to have Internet access and 35.2% do not have Internet access. The remaining companies did not respond to the question. We did find that 10.3% of the 1295 companies had Internet access and would be willing to use the Internet for transmission.

5. A Note About Security

Even if individuals and establishments are willing to respond to surveys over the Internet, there are serious risks to the confidentiality of that data. Because the Internet is essentially a public channel, data passing across the wire is subject to eavesdropping. Further, a computer connected to the Internet is subject to attack by persons who may wish to gain access to the machine for unscrupulous purposes.

The science of cryptography gives us methods to address the privacy issue. See Reference [1] for further discussion. In practice, these methods must be implemented by the developers of application software. The question of which methods are used by which software products creates an implementation problem. Fortunately, the desire of business to open up the WWW for commerce has begun to drive the development of open standards in this area. If developers meet the increasing demand for privacy in network communications, we can expect such features to become common in network application software.

Cryptography is not the whole answer. The systems on which survey data is stored must be secured to prevent unauthorized access and tampering. Careful monitoring and controls must be in place to allow sufficient access for data to be received and processed while at the same time preventing unauthorized access.

Just what is necessary to ensure that the appropriate level of security is maintained is still under review at the Census Bureau. As an agency collecting data under Title 13, the Census Bureau is stringently bound by confidentiality requirements. The success of the Census Bureau is in large part due to the public's perception that the Bureau does the utmost to protect the information that it collects. Ultimately, public perceptions about the security of the Internet may be a more significant barrier to participation than actual threats to security.

6. Approaches to an Internet Questionnaire

The Census Bureau has been researching various approaches for implementing CSAQs, both in a network and stand-alone context. To ensure that any given approach provides sufficient functionality, the CASIC office of the Census Bureau formulated the following criteria as a means of comparing options. These criteria include capabilities present with paper and pencil questionnaires and some features unique to automated questionnaires. In general, a CSAQ should have:

1. the ability to import data from local/remote databases
2. branching capability within the questionnaire
3. the ability to perform interactive edits (consistency and range)
4. secure communication
5. help features (general and context sensitive)
6. printing capability
7. exit/re-entry with retained data

The next section describes some of the methods being explored, and some of the advantages and disadvantages associated with them.

6.1 HTML Based Methods

The WWW is basically a collection of information servers ("web servers") connected to the Internet. Web servers listen and respond to requests for information. These requests come from information clients ("web browsers"). Essentially, the web browser says to the web server, "Hey! GET me the file called wonderful.html!", and the server, after a brief preamble, sends the file to the browser which made the request. The browser then displays the file or performs some other action.

While today's web browsers are capable of displaying a wide variety of file types, they were all designed to display documents written in the Hypertext Markup Language (HTML). An HTML file is a plain text file that contains embedded tags, usually enclosed in brackets. These tags identify properties of the document text to which the tags apply.

Certain HTML tags allow a document author to create "containers" in which someone viewing the document can enter information. The result is an HTML **form**, consisting of zero or more **form elements**

enclosed with a form tag (<FORM> </FORM>). Available form element types include buttons, text entry fields, checkboxes, selection lists, and radio buttons. Each of these elements possess different properties, but all of them accept user input. The user enters data to the form by performing mouse or keyboard actions ("**events**") - clicking on a button, entering text into a text box, selecting an item from a pull-down list, etc. The form elements then retain these values.

In order to use HTML forms to collect data, the data must be transmitted back to the server. Web servers support the Common Gateway Interface (CGI), which enables the server to receive information from the browser and direct it to a process (a cgi "script") that handles the data. The user "submits" the form using a special form element - the "submit" button. The data is then passed to the server (often referred to as POSTing the form) and processed by the cgi-script, which can then send messages back to the client such as a confirmation, another form, the results of edits, etc.

With HTML forms with CGI scripts described above, it is difficult to provide some of the features desirable for CSAQs, such as interactive editing and branching. Since all of the "intelligence" resides on the server, a form must first be submitted before any validation can occur.

One of the more interesting developments in web technology has been the idea of "downloadable code" - being able to write custom applications that can be downloaded with web pages to provide functionality (e.g., validate form input and implement branching) beyond what is available with HTML. The Java language, developed by Sun Microsystems, was the first language to demonstrate this. MicroSoft Corporation is developing a similar tool called ActiveX. Java is a full featured programming language similar to C++. With Java, one can create "applets" - custom applications that can be embedded within HTML documents. When a page containing an applet is loaded by a "Java enabled" browser, the applet is displayed on a portion of the page known as its frame.

JavaScript, developed by Netscape, is another language that can provide web pages with additional functionality. Like Java, JavaScript is object-oriented. However, JavaScript works differently than Java. Java applets handle events (i.e., entering text into a text box, or clicking on a button) occurring within the applet frame, allowing the programmer to create interactive applications. JavaScript can handle events anywhere in an HTML document.

HTML and CGI, together with languages like Java and JavaScript, enable the CSAQ author to meet most of the requirements described earlier. Prior period data and completed/partially completed questionnaires could be obtained from the server (if security requirements can be accommodated). Automated skip patterns, interactive edits, and help features can all be implemented with these tools. Importing data from **local** files is a problem, since both Java (in the "applet" context) and JavaScript are prohibited from accessing local filesystems. It would be convenient for the browser and web server to provide for secure data transmission, though applets could have encryption methods written into them.

What is attractive about HTML-based methods is that the user need only have a web browser which implements the features used in the CSAQ. This requirement can be satisfied if all browsers implement the necessary features, or if a few browsers meeting this criteria dominate the market. Using HTML based methods also frees the survey organization from having to develop and maintain separate versions of the CSAQ for various hardware/operating system configurations, although in practice, specific problems will probably still be associated with particular systems.

6.2. Helper Applications and Plug-ins

Anyone who has used a web browser is probably familiar with helper applications. When responding to a request for a file, a web server will send information to the browser regarding the requested file's type, for example - text/html. Based on this information, the browser can decide whether to display the file itself, or whether to pass the data to an application capable of displaying the data. Most browsers are capable of displaying text and a few image types. When a browser encounters a file of a type it does not know how to display (for example, a spreadsheet file from an accounting program) it can launch a "helper application" (the spreadsheet program) that is capable of displaying the document.

Plug-ins do the same thing, but are more tightly integrated with the browser. Plug-ins give the browser the ability to display additional file types within the browser window. While they add functionality to the browser, they do not necessarily function independently.

Many commercially available applications for designing electronic forms store the definition of the created form in a proprietary file format. The questionnaire is then administered by running an application which interprets the form definition file and "plays back" the questionnaire to the respondent, saving the respondent's input to a file. If a respondent's web browser is configured to use this survey "player" software as a helper application, links to the form definition files can be embedded in HTML documents and served over the Web. Manufacturers of electronic

forms software who implement the "player" as a plug-in can display the survey in the browser window, retaining the look and feel of the browser interface. Such plug-ins can also be written to send the survey data back to the web server via the browser.

The advantages of being able to use software that was designed for creating electronic forms are obvious. Often electronic form packages have easy authoring systems which implement branching, consistency edits, and the ability to import data from other applications. The most obvious disadvantage is that the user must have the player application, which may only be available for a price. Companies who wish to take advantage of the potential for collecting data via the Web may make their player application available for free, however the respondent must still obtain, install, and configure it for use. Companies that do make the player component available free of charge may attempt to defer the loss of revenue by increasing the cost of the authoring software.

Because this method is not tied to the browser as HTML forms are, the same form definition files that are available via the Web could be distributed non-electronically on diskettes. This eliminates the need to develop one version for the Web and another for respondents not able to use a browser. Support becomes a factor as well though. Unless the forms software is perfect, bugs that arise would likely be beyond the control of the survey agency to fix.

How you "personalize" a survey with prior period data, and security are important implementation issues. The electronic forms software might rely on security features built into the browser or implement it's own security mechanisms.

6.3 Stand-alone CSAQ

Of course, one could always simply develop a "stand-alone" CSAQ using some high level programming language. The CSAQ software could then be made available to the respondent electronically (by e-mail or FTP), or mailed on diskette or CD-ROM via USPS for those who do not have an Internet connection. Typically such stand-alone CSAQs can be programmed with edits, importing data from respondent files, branching, and printing capabilities. Once completed, encrypted data could be sent via modem, over the network, or saved to a diskette and returned by mail.

With "stand alone" CSAQ's there is no license fee per user, only development and administrative costs (support, distribution, etc.). With CSAQ's distributed as compiled binaries, the user need not possess any special software in order to use the CSAQ (e.g., a browser), however the survey agency would need to maintain different versions for different platforms. Authoring a CSAQ in a cross-platform language like Java would be ideal when (and if) Java runtime interpreters become common on all systems.

7. Potential Uses for Internet Questionnaires

Since Internet connections are more prevalent in businesses than in households, the Census Bureau is focusing on developing Internet CSAQs for economic surveys, and possibly economic censuses. There has been some prototyping and plans for prototyping Internet CSAQs. The Internet is growing, but there is still a large number of companies with access to PCs, but not to the Internet. Whether or not the Census should support diskette CSAQs in addition to Internet CSAQs for these respondents is an issue. And if we are to support diskette CSAQs, also of concern is whether diskette and Internet CSAQs should be developed using the same software. The following discussion summarizes the plans and progress made thus far. In addition, the Census Bureau is discussing an electronic means of enumeration for the Census 2000.

7.1 Case Study: Export Form Prototype

The Foreign Trade Division (FTD) of the Census Bureau collects, compiles and publishes statistical data on imports and exports. This data collection effort is accomplished in conjunction with U.S. Customs Service (Customs). Import data is collected by the U.S. Customs from import brokers. About 98% of the import transactions are filed to Customs electronically via mainframe batch files and transmitted modem to modem. Data are extracted weekly and sent on a tape from the Customs and delivered to the Census Bureau. The remaining 2% of import transactions are reported on paper. This amounts to approximately 40,000 paper documents sent to Customs and then forwarded to the Census processing office in Jeffersonville, IN.

To date, the percentage of export data received electronically is not as high as the percentage of import data. Thirty-five percent of all U.S. exports go to Canada. Because import data is often of better quality than export data, the U.S. Census Bureau electronically receives Canada's U.S. import data via a T1 line and uses this data as part of the U.S.'s export data. In 1969, an electronic system was started to collect export data. Now, approximately 300 companies report about 25% of the export transactions directly to the Census Bureau via a floppy disk, tape, or modem. The remaining 40% of export transactions are reported on paper using the Shipper's Export Declaration (SED). This translates into approximately 500,000 pieces of paper which are processed monthly. While brokers handle this reporting task for import transactions, a combination of sources report export transactions. Exporters (manufacturers) can file directly to a carrier. Often an agent or freight forwarder files the SED with the carrier. The carrier

then provides the SED and manifests to Customs. The SEDs are forwarded to Jeffersonville for processing.

Beginning in July 1995, the Census Bureau in conjunction with Customs developed an Automated Export System (AES) to replace the electronic system described above. In this system respondents report directly to Customs via modem. Data is then transmitted to the Census. At a trade conference, FTD staff were shown an HTML version of the SED form developed by a private company. This private company was a certified AES participant. Census staff from FTD then entered into contract with this company to test an HTML SED form. Census predicts that an HTML SED form would expand participation of AES for small and medium reporters. Within this contract, Census commented on the format of the HTML SED form.

By August of 1996, an HTML SED form will have been pilot tested with approximately 20 companies. To obtain 20 companies, a press release was issued on the Census Bureau's Web site. Customs faxed the release to trade members in all vessel ports. The Census Bureau received about 25 calls/faxes from this group who read the release and were interested. Approximately 18 of this group are in the process of signing up to participate. Some interested companies are working with the private firm to acquire browser capability and Internet access. Most interested companies had that capability. The Census Bureau (FTD) called about 20 small exporters (50 or fewer export shipments a month) and inquired about their interest in reporting in such a fashion. Only 2 of those companies were interested. About 50 more small companies were just faxed the release. Very few from that group responded.

The HTML SED form (see http://www.tradecompass.com/flagship/demo_frame.html for a demo form) is similar to the paper form, but it does take advantage of functions unique to electronic questionnaires. For example, there is a go to bar at the bottom of the screen, which automatically skips you to sections within the form. Help screens, pick lists, and edits are available. Once the respondent has completed it, they submit it to the web server of the certified AES provider (the software company). (A certified provider is certified to file with Customs.) The Title 13 data is encrypted using Netscape's 2.0 encrypted channel. Once on the provider's node, the data is then swept behind the firewall of the certified provider. Transmissions are batched together and sent via modem to Customs.

By the end of 1999, the old electronic system currently used by 300 companies will be discontinued. These companies will be able to file using paper, direct transmission to Customs, or if the HTML prototype is satisfactory, by Internet. Currently the demo is free, but if it becomes a means of reporting for AES, there most likely will be a fee per use from the software developer. With the prototype, the 20 companies will also have to file a paper form. If the prototype is working correctly, by August, companies will be able to discontinue the paper filing and report only with the HTML SED form. In August the Census Bureau will issue a report on this prototype.

7.2 Other Economic Area Surveys and Censuses

For the 1996 Industrial R&D Survey, which will be conducted in early 1997, the Census Bureau plans to prototype two types of Internet CSAQ and two types of diskette CSAQs. Initially all 2,800 of the 1996 R&D sample companies with known R&D expenditures over 1 million dollars will be matched to the companies who responded to the 1995 COS voluntary screener described in Section 4 of this paper. We are hoping that this match will identify approximately 50 companies having access to the WWW and willing to use it for data reporting. For companies that meet these criteria, half will be notified that they can complete the 1996 R&D survey via a form on the Web which uses a helper application called JetForm. The JetForm Filler (the application which "displays" JetForm forms) will be available as a helper application (or as a Netscape plug-in) that they can download. Once they download the JetForm Filler, users will be able to request the form from a web page and complete it. The other half will be notified that the R&D form is available as a HTML form with JavaScript. Section 6.1 describes JavaScript.

The remaining companies, not notified about an Internet CSAQ, will receive both a paper R&D questionnaire and a diskette CSAQ with the identical questions. This Windows compliant diskette CSAQ will be programmed in Delphi by the Washington Publishing Company (WPC). Respondents will have the option of submitting the survey data by diskette, modem or paper. Some of the first few companies to complete the WPC CSAQ and return it will be called and asked if they would be willing to complete the same survey but using a different diskette CSAQ. The Census Bureau's tentative plans are to create a second diskette CSAQ with JetForm BizForm. The BizForm is a version of JetForm Filler with fewer features. These companies will not need an Internet connection. The JetForm questionnaire, which would accompany the BizForm, would have the same questions as the earlier diskette form from WPC. This second form completion would be voluntary in nature. Respondents would be given an evaluation form for commenting and critiquing the two different diskette CSAQ applications.

7.3 Decennial Census Electronic Form

For Census 2000, the Census Bureau is committed to researching enumeration via an electronic

form. The question content of this electronic form would at a minimum contain the same as those on the "Be Counted" paper form used in the 1995 Census Test. During the 1995 Census Test, persons who didn't think they were counted (e.g., they didn't get a form mailed to their house) could easily enumerate themselves by picking up one of these forms at a centralized facility (e.g., libraries and post offices). The "Be Counted" form requested demographic short form data as mandated by the Census. In addition, the "Be Counted" form requested address information because this form was not mailed. For coverage purposes, it is necessary for the Census Bureau to place every person at an geocodable address on Census Day. An electronic Census form would contain at minimum the short form demographic questions and enough address information to be able to geocode the persons rostered on the form to a Census identification number. This forms centralized distribution facility would be the Internet. No previous data would be loaded into the form prior to the respondent completing it, but there would be necessity for secure communication of the data from the respondent back to the Census Bureau. Issues of form type (long and short electronic forms), data confidentiality and accuracy, as well as of telecommunications capacity at peak periods are still under investigation. For future updates see http://www.census.gov/dmd/www/2000form.html.

7.4 Collection of Data for Governments Division

Governments Division of the Census Bureau has for some time now been collecting data over the Internet. Many of Government Division's respondents (state and local government offices, libraries, universities) have access to the Web and most of the data collected from these responding units does not fall under the data confidentiality mandate, Title 13. Much of the data collection is file based. Respondents who participate, already have their data in an electronic form. They've found that putting this data into an electronic data collection FORM is tedious and involves an extra step of moving it from the file (or record) to the form. The Census Bureau offers the responding agencies with the edit checks used at headquarters. Responding agencies can also pull the instructions from Census' Web site. Of those that send files, some extract from their own databases and some key forms which they collect themselves. Data files are sent/transferred to the Census Bureau in a specified format(s). The goal is that once the Census Bureau receives a file, the data should automatically load into a database (where prior year data also resides). The data should then be edited and feedback provided to the respondent in a timely manner. Right now, Governments Division is in early stages of experimenting with automatic database loads. Some of the respondents who send files are already collecting electronic questionnaires via the Web from individual institutions, and with great success. (One example is the State of Maryland public library system.) Governments Division considers themselves fortunate in that most of their respondents are very anxious to use the Web for data submission and editing.

8. Conclusions

Although Governments Division plans to continue to use the Internet to collect data, results from the upcoming prototype tests will help to determine how or if Internet data collection will proceed in a production mode in other areas of the Census Bureau. Assuming that some sort of Internet data collection will ultimately occur, there are integration and management issues to address when multiple data collection modes are used in one survey or census. Likewise, the effect on the data when using the Internet as a data collection mode also needs to be studied. Further, the questions of how to implement a particular solution can never be answered "once and for all" since the technology keeps changing, making new tools available, and rendering old ones obsolete.

9. Acknowledgments

The authors would like to thank Charles Woods, Diane Schapira, Barbara Sedivi, Al Paez, Karl K Kindel, Kathleen E Chamberlain, and Ruth Detlefsen all of the Census Bureau for their help with this paper.

10. References

[1] Cain, Adam (1996), "Introduction to Web Security. An Overview of Technologies for Security, Authentication, and Privacy on the World-Wide Web." National Center for Supercomputing Applications. URL: http://www.ncsa.uiuc.edu/People/acain

[2] Groenfeld, Tom (1995), "The Online Safety Net." INFORMATIONWEEK, January 30, 1995.

[3] Ogden Government Services (1993), U.S. Bureau of the Census Technology Assessment of Data Collection Technologies for the Year 2000: Final Technology Assessment Report. A report prepared for the Bureau of the Census Year 2000 Staff.

[4] Oppermann, Martin (1995), "E-mail Surveys-Potentials and Pitfalls." Market Research, Summer 1995, Vol. 7 No 3, pp. 29-33.

[5] Quarterman, John S. (1994), "Preliminary Results of the Second TIC/MIDS Internet Demographic Survey." Matrix News, 4 (12), December 1994. E-Mail: mids@tic.com, http://www.tic.com

[6] Sweet, Elizabeth and Ramos, Magdalena (1995), "Evaluation Results from a Pilot Test of a Computerized Self-Administered Questionnaire (CSAQ) for the 1994 Industrial Research and Development (R&D) Survey." Economic Statistical Methods Report Series ESM-9503, U.S. Department of Commerce, Bureau of the Census.