Distr.
GENERAL

CES/AC.71/2004/10
4 March 2004

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE (ECE) CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS)
(Geneva, 17-19 May 2004)

Topic (i): Web technology in statistical information systems

INTERNET DATA COLLECTION AT THE U.S. CENSUS BUREAU THE CENSUS TAKER SYSTEM

Supporting Paper

Submitted by the U.S. Census Bureau, United States[1]

I. INTRODUCTION

1. The US Census Bureau's Mission is to serve as the leading source of quality data about the nation's people and economy. Data collection is the first step in the fulfillment of this mission. Though postal mail, personal interviewing, and telephone interviewing are the most common methods for collection, the World Wide Web (Web) offers an efficient and popular alternative collection mode. The popularity, pervasiveness and convenience of the Web have sparked a demand and expectation for Web-based alternatives to respond to questionnaires.

2. Respondents have indicated the desire for a Web reporting option. This demand has been indicated through screener questionnaires, Web evaluation questionnaire response, telephone calls, company visits, and email messages. By giving attention to the public needs and concerns, we improve the public's perception of the Census Bureau and increase public cooperation with our censuses and surveys.[2] Indeed, it is a strategic objective of the Census Bureau to reduce the reporting burden its censuses and surveys impose on the people and businesses of the U.S. The Web offers a good solution to meet this objective.

---

[1] Prepared by David Raszewski and Giuseppe Mistichelli (david.raszewski; giuseppe.mistichelli@census.gov).
[2] U. S. Census Bureau *The Evolution of Web Collection at the U. S. Census Bureau – From Research to Production.* Barbara Sedivi Gaul. Released 2001.

3. Federal mandates such as the Paperwork Reduction Act of 1995, Government Paperwork Elimination Act (GPEA) of 1998 and the E-Government Act of 2002 are also promoting the use of Web-based data collection. The Paperwork Reduction Act seeks to "*Minimize the paperwork burden for individuals, small businesses, educational and nonprofit institutions, Federal contractors, State, local, and tribal governments ... resulting from the collection of information by or for the Federal Government.*" In addition, it states that with respect to the collection of information, agencies must "*To the maximum extent practicable, use information technology to reduce burden and improve data quality, agency efficiency and responsiveness to the public.*" The GPEA requires Federal agencies to allow individuals or entities that deal with the agencies the option to submit information and to maintain records electronically, when feasible.

4. The E-Government Act, which became effective April 2003, also impacts the growing use of Web technology in the interactions of the Census Bureau and respondents. E-Government uses improved Internet-based technology to make it easy for citizens and businesses to interact with the government, save taxpayer dollars, and streamline citizen-to-government communications. While not addressing Web data collections per se, it mandates that Information Technology systems meet privacy and confidentiality standards. As such, it helps to assuage the valid concerns respondents have about submitting their data over the Web.

5. As a result of the factors above, The Census Taker Systems (CTS) was developed. Designed specifically for Internet Data Collection, CTS seeks to meet the growing demand for Web-based collection alternatives, reduce respondent burden in a user-friendly manner, and satisfy various federal mandates. CTS provides a high quality, secure and cost effective means of electronic data collection that is capable of accommodating the wide variety of surveys and censuses conducted by the Bureau. To accomplish this in an efficient manner, a standardized approach to security, architecture/hosting, programming and form design was necessary. This paper will detail the CTS application and highlight the advantages gained through the standardized, modular approach it uses for Web data collection.

Background

6. Research on the use of the Web for data collection began at the Census Bureau in the Computer Assisted Survey Research Office (CASRO). CASRO was formed in 1992 with the mission to research and begin implementation of computer assisted survey data collection methods, the goal being to identify technology that aided in the collection and processing of data in an expeditious, coordinated, and cost effective manner. CASRO began testing and implementing Web Computerized Self-Administered Questionnaires (CSAQs) in 1996 and continued through 2001. By 2001, CASRO was conducting over 15 custom designed surveys. Their research revealed the following benefits of web data collection:[3]

7. Through the CASRO experience and work done in various Bureau offices, two very different types of Web CSAQs evolved consistent with the needs of the respondent and the particular data collection effort. The first type is an interactive (on-line) form that respondents access simply by entering a specified Uniform Resource Locator (URL) in their Web browsers. The second type is an (off-line) executable program respondent's access by linking to a specified URL, downloading an installation file and installing the software on their system.

[3] U.S. Census Bureau Web Computerized Self-Administered Questionnaire Transition Plan. Computer Assisted Survey Research Office. November 29, 2001

Table 1: Web CSAQ Benefits

Better Quality Data	• Resolve keying error while reporting • Resolve/explain data anomalies while reporting • Allows researchers to bridge interactivity gap between paper and interviewed questionnaires • Respondent perception of survey automation advantageous to reporting (less time consuming, fun, etc.)
Quicker Survey Timing	• Electronic data transfer in seconds vs. postal delivery • Eliminate/reduce processing steps at Census such as keying verification, telephone follow-ups, etc.
Reduced Respondent Burden	• Automatic data fills and calculations
	• Reduce time needed to address telephone follow-up calls • Automatic skipping of non-applicable questions • Importing bulk data from spreadsheets or databases
Cost Savings to the Census Bureau	• Eliminate forms printing and storage costs • Reduce mail package preparation and postal charges • Eliminate costs associated with diskette creation and mailing • Reduce data anomalies telephone follow-up calls • Eliminate data keying and verification costs

8. As a research office, the drift into high-level production activities was not an ideal situation for CASRO or the Census Bureau. Consequently, it was decided that a transition to a full production system was needed. During this time, another Web CSAQ effort at the Census Bureau was underway aiming at the creation of a web version of the Census 2000 short form. The system was known as "Census Taker" and was created specifically for the Decennial Census. It was a customized application that consisted of a one-page Hypertext Markup (HTML) form. The system went live with the Decennial Census on April 1st 2000. Given the novelty of the Web as a collection medium, the Census Bureau was cautious and did not highly publicize the availability of this option; yet over 80,000 respondents found and used the system to answer their questionnaire. The system showed great potential in its ease of use, security, and infrastructure.

9. After the Decennial Census it was determined that customization for various Web CSAQs was no longer resource and cost efficient. A standardized approach was needed and a production home had to be found. A decision was reached to modify and expand the CTS to handle the surveys being done by CASRO, and to conduct the 2002 Economic Census. By the launch of the Economic Census in December 2002, CTS was ready to handle electronic collections from the 3 million respondent universe of the Economic Census, and satisfy the collection needs of most surveys that had been conducted in CASRO.

II. DATA COLLECTION SERVICES

10. The CTS provides two data collection services: (1) Secure File Transfers, and (2) Interactive HTML forms.

A. Secure File Transfer Service (FTS)

11. This service provides Internet users with a secure (encrypted) means of both downloading files *from* the Census Bureau and uploading files *to* the Census Bureau. FTS is used to both distribute and receive software and data files. FTS can function in two different modes depending on the requirements of the sponsor.

12. The Manual Mode requires a respondent to manually open a web browser, go to the system's login page, enter their unique user name and password, and choose the files(s) to download or select the appropriate data file(s) from their local system to upload.

13. The Automatic Mode automates all communications between a downloaded software program that is web service enabled, and CTS. This process greatly simplifies the file transfer exchange.

14. For example, a user can download a software application from the Web and then install and start the software on their personal computer. The running software then automatically communicates directly with the FTS as required for the upload or download of appropriate file(s).

15. FTS is currently used throughout the Census Bureau for different purposes. It offers a simple and standardized means of securely exchanging files with the public over the Internet. File types can vary and include: spreadsheets, ASCII text, executables (.exe), and others as needed and defined by survey sponsors.

B. Interactive Forms Service (IFS)

16. IFS provides a standardized system for collecting survey and census information by means of encrypted web page (HTML) forms. All user supplied information is encrypted both in transport and when saved. IFS is a standardized, meta data driven service that creates and securely hosts a survey. It provides an easy mechanism for the design of the web survey and the easy inclusion of edit checks and complicated skip or branching patterns. Single or multi-pages forms can be used, and the ability to preload information exists within the system. Section V will take a closer look at the Interactive Forms service.

C. Current Activity

17. The current CTS activity is outlined in Table 2 below. It shows surveys that are either in production or in the planning/design phase of development (as of 02/2004). CTS flexibility can be seen in the variety of survey types, respondent universes and sponsoring area.

Table 2: CTS Activity

Survey	Sponsor	Service Type	Frequency	Universe
Manufactures' Shipments, Inventories, and Orders (M3)	Economic	IFS	Monthly	8,000
Quarterly Financial Report (QFR)	Economic	FTS	Quarterly	6,300
Company Organization Survey (COS)	Economic	FTS	Annual	53,000
Annual Survey of Manufactures (ASM)	Economic	FTS	Annual	3,000
2002 Economic Census	Economic	FTS	5-year Cycle	3,000,000
Quarterly Services Survey	Economic	IFS	Quarterly	5,000
Current Industrial	Economic	FTS	Quarterly	15

Report				
Boundary and Annexation Survey	Geography	IFS	Annual	10,000
Private School Survey	Demographic	IFS	2-year Cycle	37,000
Teacher Follow-Up Survey	Demographic	IFS	2-year Cycle	7,000

III. META DATA – FLEXIBLE, CONSISTENT, REUSABLE

18. CTS is metadata (parameter) driven. This means that both the Interactive Forms Service and the File Transfer Service are existing programs that do not have to be written, tested, and debugged for every new survey. System code is modularized (broken into logical sections of program code for specific purposes) and need only to be modified (tested and debugged) to add new features.

19. In most cases new features can be added by means of adding a new module of code. The new code module is then made accessible to the main programs by a new meta parameter. The addition of new parameter capabilities in this manner eliminates most programming problems associated with modifying a single monolithic program.

20. A new meta data file and the running of some setup programs is all that is required to add a new Interactive Form Service or a new File Transfer Service to CTS. Once the primary meta files are in place, other setup programs can then automatically load additional standardized meta or parameter file information, setup directories, database tables etc., as required and appropriate.

21. Advantages to Metadata:

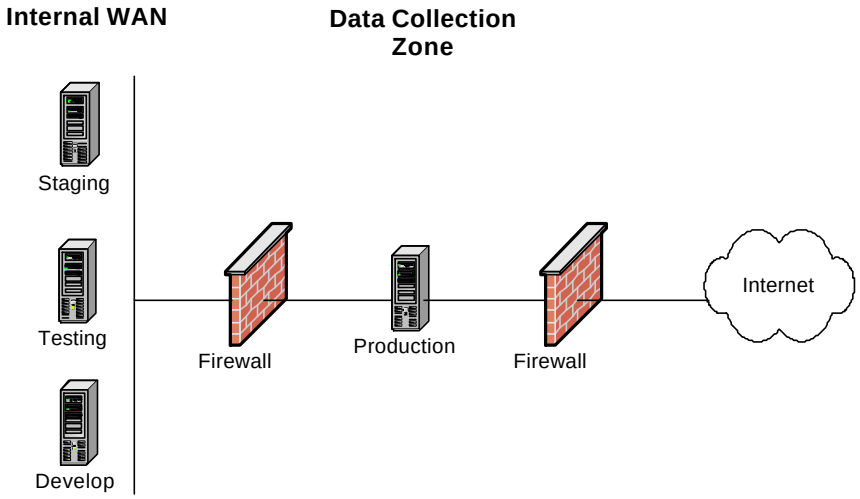
- Surveys can be designed very quickly with minimal training for survey form designers.
- Existing programs do not have to be re-written, tested, and debugged for every new survey.
- Programming knowledge is not required for most forms yet the ability to write and use complex customized subroutines is built-in.
- While the system is highly customizable, it presents a consistent look and feel to users.
- The form designer only has to worry about the layout of the questions, the skip patterns through the form, and what standardized checks are to be performed on which responses. Skip patterns through the form are much simpler to understand, construct, and check than in traditional form specification languages.

IV. INFRASTRUCTURE AND SECURITY

22. From the beginning, security considerations were of the utmost importance in the design of the CTS infrastructure. The surveys and censuses hosted by CTS collect information that is protected under federal laws. These laws guarantee not only the confidentiality of the information, but also the privacy of those reporting. The non-disclosure regulations are stringent and limit data collection activities – especially electronic data collection – many standards are to be followed and implemented.

23. As a result of security considerations, the Bureau created a separate sub-net or security zone dedicated to Internet data collection activities; CTS is the only application running in this zone. Figure 1 below illustrates a high level view of CTS infrastructure.

Figure 1: CTS Infrastructure



24. Both the National Institute of Standards and Technology (NIST) and the National Security Agency (NSA) have reviewed the basic security design for the system. All data transmissions to and from the CTS are encrypted using Secure Socket Layer (SSL). All calls to the web server use the HTTPS protocol, and only 128-bit capable browsers are allowed. Data transfers to and from the Census Bureau internal network are also encrypted through the use of Secure Shell (SSH). Along with the encryption of data transmissions, data storage on the system is encrypted using 1024-bit encryption. This data remains encrypted until it is securely transferred to the sponsor's machine on the internal network.

25. Summary of implemented security measures:

- All communications with users (including login) are encrypted.
- In order to gain access to a specific service, all users are required to:
    1. Use a browser that supports 128-bit encryption (major browsers since 1998).
    2. Provide their uniquely assigned user name and password (mailed to respondents).
- All data are encrypted or double encrypted at all times (transit and storage).
- External (from the Internet) access to the host web server is limited to HTTPS (encrypted) protocol for all interactive communications and HTTP protocol for static help and informational web pages.
- To maximize security control and further reduce communications access to the host computer, Census Taker is the only software application running on the host web server.
- Other security measures that are part of the system include intrusion detection and access logging.
- Firewalls restrict external communication (from the Internet) with Census Taker servers to only those protocols necessary. Intrusion Detection Software (IDS) is also employed.
- All communication from the Census Bureau's Wide Area Network (WAN) is restricted to Secure Shell (SSH - encrypted telnet) or http/https. No other forms of communication or other applications are allowed.
- Census Taker uses third party authentication certificates so respondents know they are securely connected to U.S. authenticated Census Bureau servers.
- Other security measures are built into the Census Taker software to automatically disable accounts, send security warnings and problem notifications to system administrators, and prevent session spoofing.
- Monitoring software is utilized and the systems configuration files are constantly checked.

- All uploads are scanned for Viruses.

26. All software and survey changes are tested and implemented on a development and testing platform before being moved to a final staging environment. From there, changes are pushed to the production environment. Data flows are always initiated from the Internal WAN, either a *push to* production or a *pull from* production.

A. Hardware

27. Below is the current hardware for the CTS.

2 Sun 280R servers
- 750 MHz cpu
- 2 gig of main memory
- 100 Gig of Raid-5 storage
- nCipher SSL Accelerator (scribe2 only)

Sun 3800 server
- 2X750 MHz CPU
- 4 Gigabytes of main memory
- 240 Gig of Sun T-3 storage (Raid-5)
- nCipher SSL Accelerator

Compaq Proliant ML530 server
- 2 1-GHz CPU processors
- 1 Gigabytes of Main Memory
- 126 Gigabytes of Raid-5 Hard Drive

B. Software

28. The CTS uses primarily open source software. The only commercial software that's used is Sun-Solaris operating system on certain machines, and the anti-virus scanning package. Not only does this help to keep cost down, but it also allows the system flexibility when considering future hardware upgrades. The main pieces of software utilized by CTS are the Apache Web Server, MySQL database, Perl Programming language and RAV anti-virus software.

V. A CLOSER LOOK AT THE INTERACTIVE FORMS SERVICE

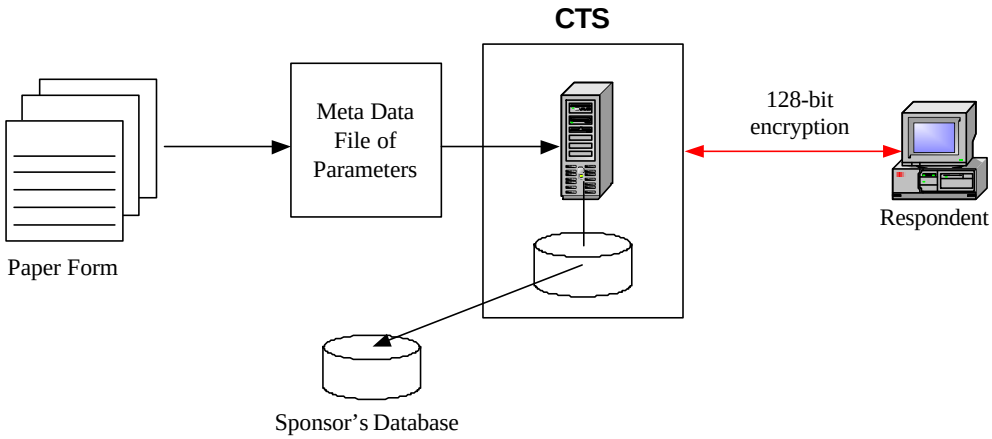
A. Forms Design and Parameter Files

29. As stated earlier, this service provides a standardized system for collecting survey and census information by means of secure (encrypted) web page forms. Forms may be long (multi-page) or short (one-page), extremely complex or simple. Forms can be designed differently and can include:

- One-page forms that scroll (often appropriate for one page paper surveys).
- Multi-page forms that have questions meaningfully grouped on a screen (semantic partitioning).
- One question or item per screen (strict item based design).
- Empty or pre-loaded with customer specific prior period data.

30. CTS allows a forms designer to easily and quickly design an interactive web survey by taking a paper form and creating a meta data file of parameters (using a simple text editor). These parameters describe the paper form in a readable language for the CTS. Figure 2 below is a high level illustration of the design process.

Figure 2: High Level Process Flow for Interactive Web Forms



31. Parameters are pre-defined and cover all aspects of the Web survey: design, branching, and data checks. Having a defined set of parameters helps ensure that across all Web surveys, high quality standards such as usability and accessibility design principles are followed; they are built-in to the CTS as defaults.

32. CTS outputs simple HTML from the parameter file definition. The screen or item designs follow a consistent visual model lending to a consistent look and feel across surveys. Major user interaction components such as the login page, main menu of options, and navigation bar are standard and consistent. Research has shown that this consistency has helped users familiarize themselves with the electronic reporting process.

B. Logic and Behavior

33. In terms of logic and behavior, simple data checks such as response data types and formats are built-in to CTS. More complex data checks are included by adding a parameter that calls customized logic components or extensions to the CTS default behavior. Branching or skip patterns are easily incorporated into a form; they are also parameter driven and for the most part require no programming knowledge.

34. Warning and error notification is clear and easy to understand. The message text is either default from the CTS or provided by the sponsor. Problems are indicated at the top of the page and at the particular question. Respondents have a choice to "ignore" the problem(s) and continue with the form. At the end of the form, a status page lists the unresolved problems in the form; a user can jump directly to each question for correction without having to navigate through the form.

35. All the behavior described above is implemented with server side programs. No client dependent scripting is used in CTS. This is done to minimize the technical problems that may arise from client-server interactions. In addition, minimizing the user's requirements helps ensure that forms hosted through the CTS can be used by the widest possible audience with the most minimal browser settings (outside of the 128-bit security requirement).

C. Context Sensitive Help and Printed Reports

36. CTS also includes a context sensitive Help option. When defining a question for a web survey, the meta file parameters (by default) create unique ID links to relevant help information. This help information is contained in a separate HTML file with unique anchors that match the ID links in the meta file.

37. Through this technique, a respondent can click on a Help icon next to a question and view the appropriate (contextual) information for that item. The help file can be simply an HTML version of instructions that already exist for the paper form, or something developed specifically for the web survey.

38. CTS also offer hard copy reports for respondents to print for their records. Reports can be in summary or full length formats. Summary format includes only the answers the respondent supplied for a question; full format lists all possible answers. In addition, there is also the option to print the form response data set to a Portable Document Format (PDF) version of the form. This technique utilizes the ADOBE® xfdf technology, which encodes form data in Extensible Markup Language (XML) for merging with the PDF file.

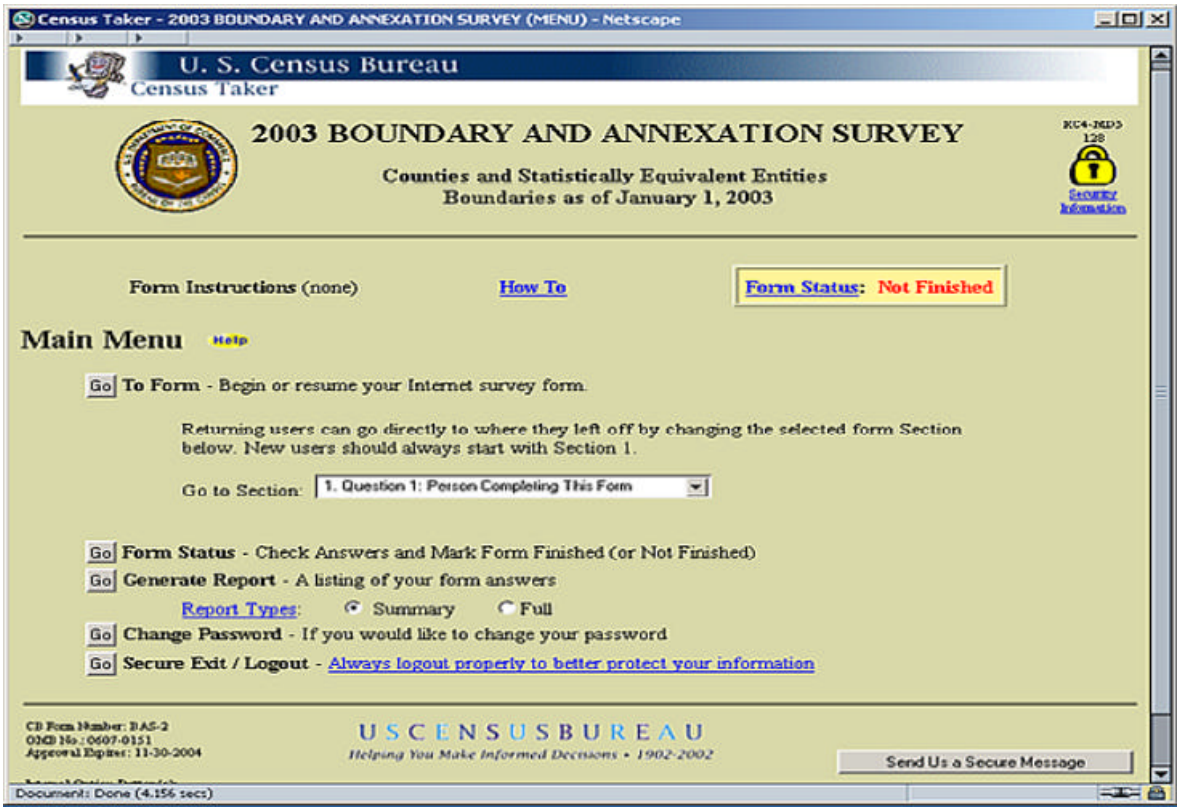
VI. CONCLUSION

39. It is clear to the U.S. Census Bureau that the use of Internet technology will increase as an alternative mode of data collection - respondents expect it and federal law mandates it. To meet this demand, the standardized, metadata driven approach taken by CTS is the most viable and sustainable framework from a corporate perspective. The CTS modular design coupled with its meta-data framework allows for new features to be included as Internet technology evolves and reach stable industry-wide usage. Examples of future enhancements include greater incorporation of such technologies as XML and Web Services.

40. In addition to application and software modifications and revisions, major hardware upgrades will be necessary within the next 12 to 18 months in order meet anticipated workload. Currently various hardware options that allow for flexibility and expansion are being investigate and include; blade server technology, clustering, and use of a storage area network (SAN).

VII. SCREEN SHOT OF CENSUS TAKER SYSTEM

41. Boundary and Annexation Survey main page.



- - - - -