

Distr.
GENERAL

CES/SEM.47/4
1 February 2002

ENGLISH ONLY

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE
EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint UNECE/Eurostat Seminar on Integrated Statistical
Information Systems and Related Matters (ISIS 2002)**
(17-19 April 2002, Geneva, Switzerland)

Topic I: Application of web technology to integrate statistics

DATA COLLECTION THROUGH WEB-BASED TECHNOLOGY

Invited paper

Submitted by the United States Bureau of the Census¹

I. INTRODUCTION

1. Who would have thought a decade ago that the Internet and Web-based technology would have the impact on our lives that it has today? This technology explosion changed not only the way we live our lives but also the way we conduct our business. At the United States Census Bureau, the emergence of the World Wide Web has changed and continues to change the way we do business as a large statistical organization.

2. The Census Bureau uses the Internet both internally and externally. Internally, the Intranet allows effective communication and transmittal of information between and among staff in both centralized and decentralized locations. Rapid innovations in Web and client/server technologies have enabled the use of the Internet as an external statistical programs tool, owing to its timely and efficient data dissemination capabilities. We aggressively pursued this opportunity and established Web sites that deal with our organization and products. These sites are accessed by data users and other interested parties throughout our country and, for that matter, the world. Most recently, the Internet has offered us a vehicle for collecting statistical information electronically. Needless to say, the Census Bureau Information Technology organization, like many others, has been challenged to keep pace with what seems an ever-changing technological, Web-based landscape, charged with innovation and opportunity.

3. This paper, then, will address the use of Web-based technology as a data capture tool. It offers to share with the reader, the U.S. Census Bureau's experience with Web-based data collection in its economic and demographic statistical programs. It will address our experience with design and implementation techniques, component-based Internet data collection software, and our recent, on-going effort to create a generic systems framework. It will also discuss the use of Web-based data collection in our recently completed 2000 Census of Population and Housing. Finally, the paper will address unique, Web-based technical issues that we encountered during the developmental process.

¹ Prepared by Richard W. Swartz (richard.w.swartz@cmail.census.gov).

II. BACKGROUND

4. Some 6 or 7 years ago, the Census Bureau concluded that there were distinct benefits to be realized by using the Internet to collect statistical information. Four such benefits were identified. First is **improved data quality**, brought about by prompt resolution of data entry errors, virtually interactive editing, and help screens. Second, **improved survey timing**, owing to the electronic movement of data as opposed to the use of conventional mail delivery systems. The reduction of processing steps such as data entry-verification and telephone follow-up for data clarification or correction also improved survey timing. Third, **reduced respondent reporting burden**, as a result of automatic questionnaire skip patterns, bulk data importing from other systems, and automatic data fills and calculations. The fourth, **cost savings**, accrue by the elimination of a number of cost centers such as forms printing, mail preparation and related postal costs, and centralized data entry and verification, to name a few.

5. Superimposed on these benefits was the growing use of Web-based communications as a way of institutional life. The emerging popularity of the Internet in the U.S., as in many countries, further supported its use as a data capture opportunity. Also, the Census Bureau and all other U.S. Federal agencies were directed, by law, to minimize the paperwork burden placed on individuals and businesses by Government. We were instructed to use information technology to not only accomplish that end but also to improve data quality.

6. A research initiative was undertaken to develop a Web-based data collection methodology and a special group of computer scientists was assembled and assigned the task. Demand for Web-based data collection, however, was growing in intensity and the research initiative quickly took on the responsibilities of a production initiative. One could say that the tail quickly began wagging the dog in our early, Web-based data collection efforts.

7. Since those earlier days, the U.S. Census Bureau has participated in fifteen separate, Web-based data collection censuses and surveys. Eleven have been successfully completed and four are in production. A sizeable number of additional census and survey applications are in various stages of development. Statistical panel sizes (sampling or census universe) range from 52 units to 3.5 million units. These applications have spanned both our demographic and economic statistics programs.

8. The majority of our applications, thus far, have been associated with our economic statistics program where electronic data reporting is especially cost efficient and convenient for respondent businesses. After all, few businesses in the United States are without a computer today. Those businesses, particularly the larger ones, reap sizeable benefits from their ability to import data from their existing systems to either an electronic questionnaire or transmit company files in predetermined formats that meet data reporting requirements via the Internet. However and for the first time, the 2000 Census of Population and Housing, our premier demographic program, also made use of Web-based data collection on a limited scale. This experience is described below.

III. THE 2000 CENSUS OF POPULATION AND HOUSING EXPERIENCE

9. The first, largest, and most worrisome application of Web-based data collection in a demographic, household program came with our recently completed 2000 Census of Population and Housing.

10. As of August 2000, it was estimated that some 51% of the households in the United States had a computer. Approximately 41% of United States households had access to the Internet. It follows that more than 80% of households with a computer had Internet access. This sizeable, potential electronic universe presented census planners with a new and exciting data collection challenge and opportunity.

11. It became obvious to census planners that the 2000 Decennial Census was a candidate for an optional, Web-based data collection application. Unfortunately, this conclusion was reached rather late in the planning cycle, creating a risk factor that many found troublesome. The Census pretest period (a pre-census

period of procedural and operational testing and adjustment) and the Census dress rehearsal (a final hands-on, dry run of the census process) had passed. Originally, it had been decided to include Web-based data collection in the census dress rehearsal; however, that plan was abandoned because of security concerns. The subject was again revisited in late 1998 and it was decided to reinstate optional Web-based data collection for the 2000 Census. Needless to say, the planning and design window had grown incredibly tight!

12. The U.S. Census Bureau had never used the Internet to collect data in a previous census or demographic program. This would be a completely new experience and our first household data collection experience. There were sizeable concerns regarding security, integrity, and load capacity. We were dealing with a new, untested system that was being developed in a compressed time frame. Nevertheless, the decision was made to move forward and introduce Web-based data collection in the 2000 Decennial Census, but introduce it on a somewhat limited scale. The availability of the Internet for response to the 2000 Census was not heavily advertised. While the U.S. Census Bureau decided to make Web-based data collection available to the public, it was hesitant to encourage massive participation for the reasons enumerated in this paper.

13. As a result of timing constraints and security concerns, certain parameters had to be placed upon the system. First, it would be designed and used to collect only basic census data, that is, 100% or short form data. Two forms are used in our census, a short form and a long form. The long form, which in addition to the short form information (questions asked of 100% of the population), contains a set of sample questions covering about 16 percent of U.S. households. Those receiving the long form were not given the option of Web-based response. Second, respondents were required to use a unique 22-digit census identification code in order to access the electronic form in order to assuage security concerns. Third, the electronic version of the form would be made available to the public in the English language only. While the first and third of these parameters were not desirable, they were necessary in order to have a reliable system in place and ready on time. Also, lacking historical experience, it was impossible to predict electronic response rate, so the system was designed to handle very high volume, millions of responses on a tight time line.

14. The U.S. Census Bureau prides itself highly on the confidentiality of its data collection programs and its track record. It goes without saying that security was a major concern and consideration throughout the systems design process. Our operational premise was that electronic data capture systems had to be secure and impervious to attack. The unique 22-digit identifier helped to ensure that falsification of data and the "creation" of new, artificial households were not an issue. Further, the unique identifier prevented hackers from guessing ID numbers at random. For a given unique census identifier, the system was designed to accept only a single questionnaire submission for processing. However, there was still the chance that a hacker would resubmit forms repetitively under the same 22-digit identifier and disable the system in much the same way other Web sites have been attacked -- from overload. Anticipating this as a potential problem, the systems designers created a safeguard against attempts at overloading. If a respondent repeatedly submitted forms under the same census identifier, submissions would be accepted at an increasingly slower rate, effectively removing the respondent's ability to overburden the system. Also, once a predetermined number of repetitive submissions was received, the system was designed to alert the systems manager. These safeguards and preventive measures tested successfully. As it turned out, we experienced no security breaches or successful attacks on the system.

15. In the end, the Web-based data collection system worked beautifully, given the constraints mentioned earlier. More than 66 thousand households chose to use the Internet response mode, representing over 180 thousand people. While this represents a very small portion of the U.S. population, it is very relevant to note that the systems and underlying methodology worked flawlessly.

16. The Internet was a very viable response mode for the 2000 Decennial Census. Load testing showed that we could have handled substantial volume, an estimated 5 million responses daily. The transition from data capture to data processing was seamless and fully electronic. There were six dedicated Web servers in two separate locations to provide capacity, redundancy, and security. Data were then accumulated on a single, dedicated server for daily transport into the Bureau's data processing systems. Integration with other

data capture medium was seamless. As the reader will quickly conclude, this data collection technique virtually eliminated data capture costs, whether keying or image-based.

17. Web-based data collection has earned a prominent place in our future Decennial Census planning. It is too early to tell exactly what form it will take or what methodology will be used for our 2010 Census. For that matter, it is impossible to know exactly what we now call the "Internet" will look like several years from now. We are certain, however, that it will play a key and larger role in our next national population and housing census

IV. U.S. CENSUS BUREAU SYSTEM: CONCEPT AND DESIGN

18. As stated earlier, the U.S. Census Bureau recognized the Internet as a vast and efficient mechanism for both collecting and disseminating data about American society. We have been experimenting, developing, and prototyping Web-based data collection since 1996. Just recently, a staff was created in our centralized Information Technology (IT) organization that will manage our Web-based data collection program in a production environment with formal procedures and controls. The research aspects of this work were previously described in a paper authored by Barbara Sedivi Gaul, U.S. Bureau of the Census, and presented to the Joint UNECE/EUROSTAT work session on Electronic Data Reporting in February 2002.

19. Our conceptual goal for the production environment is to create a "framework" system that can be adapted for multiple census and survey use – a core system. Most of the underlying software needed is now being rewritten to make it more robust and the system more stable and uniform. We are attempting something of an "off the shelf" or framework approach in our design; that is, 'one size fits all'!

20. There are two basic methodologies that we use to deliver an electronic questionnaire to a survey or census respondent. They are the *downloadable-executable method* and the *interactive method*. The downloadable-executable method has two delivery options. We can deliver a completely self-contained electronic questionnaire to the respondent's computer through the Internet. This self-contained questionnaire is in the form of an executable file that will run on any personal computer with a Microsoft Windows operating system. The respondent will download the file to disc, complete the questionnaire, and then return the completed questionnaire over the Internet. This system is basically a file transfer system. The second delivery option we use is to provide the respondent with a 3-1/2 inch computer disc, mailed through our traditional postal service. When using this method, all we have basically changed is the delivery mechanism. The respondent will either complete and return the disc to the U.S. Census Bureau by the postal service or use the disc to enable a computer-generated response via the Internet.

21. The second methodology, the interactive method, is our preferred methodology. Here, we deliver the questionnaire electronically, but through a series of dynamically generated HTML Web pages. Communications are interactive while the questionnaire is being completed by the respondent. This method was used successfully during the 2000 Census of Population and Housing described earlier. This methodology is highly desirable. It is a more expedient method, provides maximum flexibility (interactive editing, data aggregation, etc.), quick set-up, standardization, and sizable operational cost savings. We have developed standards for each of these two methods and the overall concept is referred to as Computerized Self-Administered Questionnaire (CSAQ).

22. Throughout its development, there have been many concerns with Web-based questionnaire delivery and data capture. These concerns tend to be very similar to those we have encountered numerous times in the past when introducing new, innovative data collection techniques. For example, when we went from the traditional paper and pencil questionnaire to laptop computer data entry or to telephone interviewing and keying for our surveys, many of the same concerns were expressed -- security, respondent bias, etc. The Internet is just another evolution in this path of innovation.

V. UNIQUE WEB-BASED TECHNICAL ISSUES

23. During systems design, we found a number of technical issues that needed special attention and four

of these stood out clearly.

- ✍✍ Internet browser inconsistencies;
- ✍✍ User authentication;
- ✍✍ Encryption;
- ✍✍ Data security.

There are two predominant browsers in the U.S. – Netscape and Internet Explorer. Each of these browsers has different levels or versions and different security options that can be set by the user. These differences make it very difficult to write a single set of software that will accommodate both browsers and their idiosyncrasies. For example, if the carriage return key is pressed using the Internet Explorer browser the electronic form will be transmitted, but the same procedure will not transmit it using the Netscape browser. The Census Bureau found that writing software to the lowest common denominator, in this case HTML, coupled with extensive laboratory testing using different screen sizes and browser levels or versions, enabled accommodation of both type browsers. Writing software that identifies browser type and version has also been very helpful.

24. Proper user identification was a great concern and posed a difficulty for us. Unfortunately, personal digital certificates for a Private Key Infrastructure (PKI) solution are not readily available in the U.S. Also, a PKI solution would be very expensive to establish. Our solution (excluding our 2000 Census) was to mail a user ID and password to respondents in two separate mailings using our traditional postal service. This solution has been successful and is low in cost.

25. As mentioned earlier, data security is paramount to us; therefore, data encryption is an important component of our systems design. We concluded that a 128-bit encryption and secure socket level (HTTPS) support for transmission of data were critical. We decided to use a global server certification that boosts the client browser to the highest level of encryption available between server and browser. At present, this is 128-bit encryption.

26. Security is further enhanced once data are transmitted to our servers. Upon receipt the data are compressed and we apply 1024-bit encryption at the record level. Should a hacker penetrate our Internet data collection system, he or she would find the collected data unintelligible. These census or survey data reside on servers protected by a firewall for a short period of time and are then moved to our internal computer systems for processing. Once data are transmitted to our internal computer systems, they are protected by our perimeter firewall, our highest level of security and protection.

VI. ORGANIZATIONAL RELATIONSHIPS DURING SYSTEMS DESIGN

27. The inception of a Web-based data collection initiative for one of our censuses or surveys is truly a collaborate effort within our organizational structure. When a statistical program manager wants to conduct a survey or census using the Internet, he or she will meet with members of our central IT Internet staff who coordinate various Internet activities. The first determination is which electronic questionnaire methodology and delivery option to use. If the downloadable-executable method (3-1/2 inch disc or electronic file transfer) is determined the best approach, the program manager will be given questionnaire examples, a copy of the Internet standards, and will be directed to a group of computer professionals in either our Economic or Demographic Program areas who write such CSAQ surveys. The delivery and coordination of the case management information is then handled between these two groups. If the interactive method (interactive Web pages) is chosen, the central IT Internet staff will assist statistical program managers in writing instructions for the electronic questionnaire and systems support. The central IT staff and project managers in statistical program area will then handle the logistics and case management of the questionnaire jointly.

VII. ROLE OF THE CENTRAL IT ORGANIZATION

28. The central IT organization in a statistical agency has a very important role in Web-based data collection. Such a data collection system has two very broad components – (1) an electronic questionnaire,

and (2) everything else associated with moving that electronic questionnaire to and from a respondent, including systems and security considerations. At the Census Bureau, the central IT organization participates in the design of the electronic questionnaire as a technical expert with statistical program managers.

29. The central IT organization provides a secure framework and uniform methodology for the project. It writes and manages standardized software, manages the applications, hardware systems, networking, and security. The statistical program managers design their Web-based program to meet their unique program needs, e.g. questionnaire design, skip patterns, edits and checks, and the like. The Web-based design is accomplished within the framework system made available by central IT.

30. Close coordination between computer and statistical professionals assures survey objectives are met. Such coordination also assures that the electronic questionnaire is compatible/acceptable to the systems that will move it from place to place and eventually process it. Once the survey instrument is designed and ready for use, the central IT organization is responsible for all remaining steps and processes. The most important of these include:

- ~~SES~~ creation, installation, and testing of Internet software;
- ~~SES~~ security systems and firewall management;
- ~~SES~~ data base administration and back-up systems;
- ~~SES~~ hardware procurement, configuration, and management;
- ~~SES~~ server software and configuration;
- ~~SES~~ Web electronic questionnaire set up;
- ~~SES~~ user account management.

While delineation of responsibilities and duties between statistical program managers and IT specialists appear crisply defined, there tends to be an ongoing dialogue throughout the entire census or survey planning, design, and implementation process.

VIII. CONCLUSION

31. In our experience to date, it has become clear that Web-based data capture has a very real application in our statistical programs, both sample surveys and censuses. Thus far, our experiences have been positive and our program applications have been increasing. Over the past decade or longer, the trend at the U.S. Census Bureau has been toward electronic movement of paper. Web-based data collection is the next logical progression of this trend, which has included decentralized data entry, lap-top computer data capture by survey takers, telephone data capture and entry, and a variety of image-based data entry techniques.

32. Our economic statistics programs have been the primary beneficiaries of this new methodology thus far. However, we are finding increasing applications in our demographic programs as well. The latter, of course, is somewhat limited since every household in the U.S. does not yet have a computer. Nevertheless, Web-based data collection certainly lends itself to voluntary participation. We can only expect to see an acceleration of this data collection methodology in the future.