WEB-BASED DATA COLLECTION IN NATIONAL SCIENCE FOUNDATION SURVEYS

Ronald L. Meeks, Ann T. Lanier, Ronald S. Fecso, National Science Foundation, Mary A. Collins, Westat, Inc.
Ronald L. Meeks, NSF, 4201 Wilson Blvd., Suite 965, Arlington, VA 22230

Key Words: CSAQ, CASIC, Internet

The National Science Foundation (NSF), Division of Science Resources Studies (SRS), conducts periodic surveys of academic institutions, Federal agencies, private corporations, and individuals to support its mission of collecting and disseminating information on science and engineering resources in the United States. Recently, NSF has been developing approaches to collecting data through the World Wide Web, to support greater flexibility and quality in its surveys. There are several technical issues involved in efforts to collect information over the web. Applying confidentiality laws is a new area, with securing of the data during the session and server protection from "hackers" needing thought. The desire to conduct surveys over the web is inspired by belief that data quality can be improved by interactive editing during the session and direct respondent corrections. Yet, other nonsampling error issues, with this new twist, include sampling frame completeness and knowing who actually filled out the instrument when several people may have access to the e-mail account. Much needs to be learned about the cognitive processes when a potential respondent "opens their mail." What is the best "look" of the form? Will it convey the same official sense as agency letterhead? And finally, will response rates be better or worse, and why? At the present time, several NSF surveys are in various stages of development in the move toward greater use of web-based data collection. This paper discusses our experiences to date.

Surveys with Web-based Data Collection

Six of the Foundation's surveys are developing or have developed web-based approaches to collecting data. Five of these studies are institutional surveys, while one is a survey of individuals. The five institutional surveys include:

- The Survey of Graduate Students and Postdoctorates, the Survey of Academic R&D Facilities, and the Survey of Academic R&D Expenditures, which collect data from colleges and universities;
- The Survey of Federal Funds for R&D, which collects information from federal agencies; and
- The Survey of Industrial R&D, which collects information from private corporations.

The National Survey of Recent College Graduates is a study of individuals who have received bachelor's or master's degrees in the sciences and engineering.

Advantages of a Web-based Approach

A web-based system can combine attractive features of the three types of data collection approaches used in these six surveys: paper questionnaires, data collection programs on diskettes, and computer-assisted telephone interviewing (CATI). As with a CATI study, web-based data collection can be used to track the status of all cases continuously. This permits timely and effective nonresponse followup. Web-based collection also allows survey administrators to easily incorporate corrections or clarifications that are identified through early responses to the survey. Thus later respondents may benefit from the experiences of each early respondent who provides feedback.

As with both diskette-based and CATI data collection, web-based instruments also support data quality to a greater extent than paper surveys because the automated instruments can include checks for inter-time and inter-item consistency, enforce allowable ranges, and control skip patterns to ensure that appropriate questions are presented to the respondent. In this way, the automated instrument can reduce respondent error, thereby reducing the need for followup on critical survey items.

Web-based data collection can also bring the savings in data preparation labor that are associated with diskette-based and CATI data collection. Far less manual coding is needed than is often required with paper surveys. Further savings over a diskette-based approach accrue with a web-based design because diskette preparation and distribution are not required. In comparison to CATI surveys, interviewer time is not required for responses completed via the web.

Like other self-administered data collection modes, web-based surveys are convenient for respondents because they can complete the survey at a time that is convenient for them. They can also begin the questionnaire, answer part of it, and return to complete it at a later time. Web-based instruments can also assist respondents in completing the questionnaire by providing online help features.

Another potential advantage over a paper and pencil questionnaire is that the web-based instrument can be programmed to allow institutions to save the data they enter as a flat file that they could share with others. Pre-

viously, institutions had to wait until NSF released the data to be able to easily compare themselves with others. With the web-based survey, they will be able to share electronic files with colleagues before NSF releases the data.

General Issues in Web-based Data Collection

There are a number of important issues that must be considered in development of a web-based system for data collection. One key issue is the need to protect the confidentiality of the data that are collected. As NSF has consulted with respondents to a number of its surveys, concerns about confidentiality have emerged repeatedly. Security measures such as password protection and encryption are important elements of system development. These features help to ensure the confidentiality of data and increase respondent confidence in responding via the web.

It is also important to recognize that not all respondents have access to the World Wide Web. Further, time spent online can cost respondents money if they do not have unlimited access arrangements. These limitations require that another avenue of response be available to those respondents who cannot or are unwilling to utilize the web-based instrument.

A key challenge in the development of a web-based data collection system is deciding upon system design and instrument features that meet the needs of the particular instrument under development and the respondent population. For example, some of the Foundation's institutional surveys collect the same or similar information over time and it is desirable to allow respondents to see their data from previous years. In some cases, the institutional surveys collect information that requires responses from multiple offices or persons within each institution. Strategies for dealing with these issues will be covered in the discussion of specific surveys.

Given the newness of the World Wide Web as a vehicle for collecting survey data on a large scale, it is important for developers to obtain feedback from users. This feedback can be used to assess the web-based response experience, to identify problem areas, and to develop appropriate solutions. Developers must also consider an assessment of available software systems, the portability of the instrument, and accessibility to users with a variety of system configurations.

Survey of Graduate Students and Postdoctorates

NSF's Survey of Graduate Students and Postdoctorates in Science and Engineering is an annual survey of the number of graduate students and postdoctoral fellows at academic institutions—722 schools provide data on 11,615 departments. The survey is, for many schools, a two-stage process. Department coordinators provide data to institutional coordinators who in turn provide data to NSF. The graduate student survey has historically used a mail questionnaire approach in which respondents completed a paper instrument. The project is now in the process of moving to a web-based application.

In June 1997, NSF held a workshop with a number of survey participants to demonstrate the prototype of the web-based instrument. The workshop participants provided extensive feedback to NSF regarding the desired features of a web-based application. Issues and suggestions raised by the workshop participants included:

- The need for a comments section so that respondents could document how they derived the data;
- The need for a sort feature, so that coordinators could search programs or departments by identification number or name;
- The ability to view data from prior years;
- The desirability of a spreadsheet option, so that school coordinators could send the spreadsheet to department coordinators;
- Pre-filling of fields;
- Having an online tutorial;
- Being able to submit data (and receive it back) in a flat file, rather than keying it in to web screen; and
- Data security issues.

Security issues were of particular concern to participants. Among the specific issues discussed were the need to have a password to protect schools' data and the need to have separate passwords for departments. Another issue raised was the desirability of having multiple "locks" in the instrument so that separate parts of the survey could be locked individually if desired.

The workshop participants as well as one other institution (9 schools and 400 departments altogether) agreed to participate in a beta test in the spring of 1998. As of May 1, six of these test participants had submitted data. Three of the responding institutions uploaded a flat file, rather than keying data in the web-based application.

Beta testers were able to send electronic mail messages (e-mails) as they tested the system and a follow-up e-mail survey was sent to all testers after they completed and submitted the survey. All of the respondents preferred the web version to the paper version. Responses indicated that most participants took longer to complete the web survey than they had in the past to complete the paper version. This had to do in part with testing a new system, and in part with slow system response times. Most respondents desired a way to print all of the screens with one command rather than printing screens individually. Most also wanted a reduction in the number of screens as well as a means of simplifying the handling of zeroes.

Once revisions to the system are made as a result of the beta test, the web-based data collection system will be available to the full survey population for the fall 1998 survey.

Survey of Academic R&D Expenditures

The annual Survey of R&D Expenditures is conducted with 500 to 700 research-performingacademic institutions. Since the late 1980s, the survey has used a diskette-based data collection approach. In addition, respondents have had the opportunity to respond using a file transfer protocol (or FTP) or by sending their responses through e-mail. The survey will move to web-based data collection with the FY 1998 survey.

A development workshop was held to solicit input from potential users; many of the respondents' suggestions and comments about usability and data security were similar to those already discussed above. The R&D expenditures system was beta–tested in the spring and summer of 1998.

A number of security features are built into the web-based data collection system. Each respondent receives a unique institutional ID, as well as an individual password. The expenditures survey is one that may have multiple respondents within institutions. In addition to multiple-person access to the web-based instrument, another way to handle multi-level access is the off-line spreadsheet option. The respondent downloads the sample spreadsheet and enters data using Excel. The spreadsheet can be made read-only for other staff, using Excel's password protection function. The respondent can use all Excel features, including the option to download previous years' data as spreadsheets.

Data are entered in three ways in the web-based forms, including text boxes, radio buttons, and pick lists. Error checking has been built into the system. The error messages that appear on the computer screen feature their own online help. If an error message appears and the respondent does not know what to do, clicking on the message will bring up more help on the screen. A text box at the bottom of the Web-based version will allow respondents to type notes about how the data were compiled and comment on any major deviations from previous years. At any time, respondents can print out the Web page, including any data entered to date. The data can always be saved and changed prior to final submittal. The user indicates when the data are being submitted as final. The password system ensures that only the survey coordinator can submit the data.

In November 1998, the web-based data collection system will be made available to the full survey population. A paper version will also be distributed. The diskette-based Automated Survey Questionnaire (ASQ), currently used by about 55 percent of the respondents, will no longer be distributed.

Academic R&D Facilities Survey

The Academic R&D Facilities Survey is a biennial study of 365 research-performing higher education institutions. The target population consists of schools that grant graduate science and engineering degrees and/or annually perform at least $50,000 of separately budgeted research and development (R&D). This congressionally mandated study serves as the primary source of information on expenditures and needs for science and engineering research facilities within universities and 4-year colleges in the United States.

The R&D facilities study is moving from a diskette-based data collection system to a web-based system. The design of the web-based application for the facilities survey benefited from the input of an advisory panel and feedback obtained through conference presentations. As a result of these efforts, a system was developed that is compatible with all browsers. This universal compatibility maximizes the opportunity for institutions to respond via the web. Like the web-based applications for other NSF surveys described above, the R&D facilities instrument includes edit checks, security features, and so on.

The 1998 web-based survey is currently in the field. Respondents receive a paper survey, web-instructions, login information, a unique password, and a copy of their institution's 1996 responses to the survey.

Survey of Federal Funds for R&D

The Survey of Federal Funds for Research and Development is the primary source of information about federal funding for R&D in the United States. This survey of federal agencies collects information on funding levels and future obligations for research and development. In 1996, data were collected from about 30 agencies and about 100 reporting units. The great majority of data, about 90 percent, are collected using an automated diskette program.

A web-based approach to collecting these data is currently being developed. A pilot test with selected agencies and program offices is scheduled for October 1998. The planned design will permit agencies to see and update their own information (using a Microsoft Access database), but will not permit them to access data from other agencies. Edits and automatic data summaries are two of the features that will be used to support high data quality. Survey security will be maintained through the use of unique user passwords. An online help feature will be provided to assist users in accessing and completing the web-based instrument.

Survey of Industrial Research and Development

The target population of the Survey of Industrial Research and Development consists of all industrial com-

panies that perform R&D in the United States. Of these, 25,000 are surveyed annually. This survey is the primary source of information on R&D performed by industry within the United States. Government agencies, corporations, and research organizations use the data to investigate productivity determinants, formulate tax policy, assess trends in R&D expenditures, and compare individual company performance with industry averages.

The survey is conducted by the Census Bureau under an interagency agreement with SRS. Currently, survey instruments are sent and data are collected by mail. The Census Bureau is developing web-based reporting systems for several in-house surveys as well as the NSF-sponsored industry R&D survey. The NSF survey will be included in the Census-wide pilot effort (as discussed in this session in Elizabeth Nichols' paper, *Economic Data Collection Via the Web: A Census Bureau Case Study).*

National Survey of Recent College Graduates

The National Survey of Recent College Graduates (NSRCG) provides information about individuals who recently obtained bachelor's or master's degrees in a science or engineering field. The survey is currently conducted primarily by computer-assisted telephone interviewing. Mail questionnaires are sent to non-respondents near the end of the data-collection period in an effort to increase response rates.

The use of web-based data collection for surveys of individuals has a serious limitation–specifically, many people do not have access to the web. However, recent science and engineering graduates seemed to be an ideal population on which to test web-based data collection in a survey of individuals. Based on their relatively recent university experience and their chosen fields of study, it was hypothesized that most of this population would be relatively computer literate and have access to the World Wide Web.

The 1997 NSRCG sample members were surveyed about web-based data collection issues. While all of the data are not yet available, preliminary results indicate that the majority of these graduates do have access to the web either at home or at work, and that a majority of them would be willing to respond to a survey like the NSRCG over the web. Among those who stated that they would not respond via the web, the chief concern was that they did not believe that the confidentiality of the responses could be guaranteed.

The NSRCG web-based instrument was designed as a "downloadable executable." That is, the instrument is an executable program that will be downloaded to the respondent's local hard drive, completed at his/her convenience, and retrieved when the respondent logs in again after completing the instrument.

The respondents will be sent an introductory letter about the survey, the web address, and an individual user identification number and password by mail. Respondents logging into the web site will complete a short set of verification questions to ensure that the material was sent to the correct person. The executable questionnaire will then be automatically downloaded to the respondent's local hard drive if he/she successfully completes the verification process. The respondent may complete the entire questionnaire at one time or may answer in sections. Skip patterns are controlled and radio buttons for responses are used when feasible to reduce key errors. When the questionnaire is completed, the respondent logs back into the web site to transmit the completed questionnaire. For security purposes, only the responses will be retrieved; the question text will be separated from the answers. For added security, the data file will be encrypted prior to retrieval.

Later this summer, a sample of graduates will be asked to test the NSRCG web-based data collection system. Following the completion of this test, Westat will make recommendations to NSF concerning the implementation of a web-based response option for the 1999 survey. Based on the preliminary data from 1997 survey respondents, a web-based survey appears to be a feasible and cost-effective approach to collecting data from college graduates.

Summary

At the present time, several National Science Foundation surveys are in the process of designing, testing, and implementing web-based data collection. This new approach will replace or supplement a variety of current data collection approaches, including paper instruments, diskette-based systems, and CATI data collection.

The web-based applications carry some of the best features of diskette-based and CATI data collection. Perhaps the most important of these features are online editing for allowable responses, elimination of the need for key-entry of survey data, and automated skip patterns. The web-based approach retains the respondent convenience associated with paper or diskette-based collection while eliminating the need to distribute and track paper questionnaires or diskettes. Unlike CATI interviews, the completion of the web-based survey does not require interviewer labor; a significant cost item in CATI surveys.

A key concern of respondents is the protection of the confidentiality of the data they provide. Each of the surveys we have discussed has taken steps to ensure data security using features such as unique user passwords and, in some cases, encryption of responses.

The system features of a web-based data collection must take into account the nature of the data being collected and the respondent population. Thus, editing features, pre-filled items, and definitions or instructions must be presented in such a way as to facilitate response.

It is important to recognize that not all survey respondents have access to the World Wide Web and some

respondents may be uncomfortable with using this response mode. As a result, data collection plans must accommodate alternative forms of response. Nonresponse followup strategies must include appropriate methods to collect responses and ensure high response rates.

The National Science Foundation is optimistic about the use of web-based data collection approaches and looks forward to the reporting on the success of these efforts in the future.