WEB-BASED SURVEY TOOLS

Sarah Nusser, Iowa State University; DeanThompson, U.S. Department of Agriculture
Sarah Nusser, Department of Statistics, Iowa State University, Ames, IA 50011-1210

Keywords: WWW, survey project management, CASIC, computer-assisted learning

Abstract

The World Wide Web provides an effective means of supporting survey projects, particularly when data collectors are geographically dispersed. Wireless and wireline communications can be used to integrate the survey team by providing current and consistent supporting materials to all members of the project team. Interactive tutorials, survey instructions and updates, technical support, computer-assisted survey instrument software and updates, text-based and graphic survey management reports, data views for monitoring and editing, and summary reports are some of the tools that can be delivered via Web browsers to data collection staff, survey managers, clients, and the public. We will describe Web-based tools that have been developed to support a national survey of natural resources, and discuss possible extensions of this work.

1 Introduction

Many large-scale surveys require coordination of multiple teams of interviewers or data gatherers. Computer-assisted survey information collection (CASIC) systems have been used effectively to manage and maintain consistency in the data collection process in such settings. An extension of this concept involves integrating Web-based tools with the CASIC paradigm to produce enhanced survey management.

The U.S. Department of Agriculture and the Iowa State University Statistical Laboratory collaborate on a large longitudinal natural resource survey which involves several hundred data gatherers at dozens of sites. Computer-assisted data collection for recent surveys has involved the use of handheld computers as a means of gathering data with a dispersed and mobile workforce. The data gatherers and survey managers use handheld computers and desktop PCs to connect to the central server for various functions, including picking up samples for data collection, returning data, obtaining software updates, accessing training materials, and viewing data, instructions, and progress reports. The purpose of this paper is to describe our experiences with developing Web-based tools to support various phases of sample surveys.

We begin by describing the survey setting that motivated our research and related settings in human population surveys. We then review Web-based tools that have been developed to support a large national survey and outline areas we expect to explore in future research.

2 National Resources Inventory Surveys

Our original objective was to develop CASIC methods for a large survey called the National Resources Inventory (NRI). The NRI is a national survey program of the U.S. Department of Agriculture designed to monitor conditions and trends for natural resources on nonfederal lands (Nusser and Goebel, 1997). Its primary purpose is to provide information for agricultural policy evaluation and development, and to support agro-environmental research objectives, such as modeling economic and environmental effects of alternative agricultural policies, or biophysical processes such as pesticide leaching. Special topic investigations are also conducted regularly as part of the NRI program, for example, to monitor the effects of recent farm legislation on conservation practices.

The Foundation NRI survey is based on a stratified two-stage area sample of U.S. lands consisting of about 300,000 primary sampling units (PSU), which are usually defined to be 160 acre (64 hectare) square area segments. Approximately 800,000 secondary sampling units, i.e., points in the PSU, are selected in the second stage of sampling. The survey is longitudinal, with data collection occurring on sample units every five years.

Data collection is based on remote sensing (primarily photo-interpretation), with effort devoted to abstraction of office records to obtain information from conservation plans and soil surveys. Field visits are made to sample units when data are not available through standard materials, or when it is required for special studies. Within the PSU, areas of polygons defining specific land uses are recorded, as well as the length of linear features such as streams. Within the primary sampling units, usually three points are selected. The variables collected at each point include land use classifications, agricultural practices, soil characteristics, and other natural resource attributes, such as habitat and wetland information.

3 Distributed Survey Settings

Data collection activities for the NRI are organized by Inventory Coordination and Collection Sites (ICCSs) located through the U.S. There are approximately 20 ICCSs, although in practice, data gatherers are housed in dozens of sites. A handheld computer is provided to each data collector. Each ICCS and most data collection sites also have access to at least one Windows-based PC and/or Unix workstation.

The survey setting for the NRI requires the data gatherer to be mobile at a number of different scales. Within an office, data gatherers need to be able to move between photo-interpretation stations, paper files, and desktop computers that provide access to GIS tools and the Web. On a larger scale of mobility, data gatherers may need to go to other agency offices to abstract information from records, or they may need to go out to the field in order to observe conditions. In addition, NRI and related surveys frequently require rapid deployment, and typically involve final deadlines that are dictated by congressional or agency needs. Materials must be readily disseminated and updated, and progress needs to be monitored weekly, if not daily, during critical phases of the study.

Many sample surveys of human populations involve similar settings. For example, PSU enumeration or face-to-face interviews for national or regional populations typically involve multiple teams of enumerators. Office record abstraction or pricing surveys frequently involve remote groups of data gatherers, and large-scale telephone operations may need to integrate activities across several computer-assisted telephone interview (CATI) labs. Even within a single CATI lab, it can be advantageous to centralize survey resources. Time constraints and the need to monitor data collection and processing in an efficient manner are also a standard feature in any sample survey.

4 CASIC System Design and the Web

The components of the NRI data collection system are analogous to those of a standard desktop or laptop computer CASIC system, in which remote computers connect to and exchange information with a central database service. For the NRI, a handheld or desktop computer is used connect via a phone line or network to a central service at Iowa State University (Nusser et al. 1996, Nusser and Thompson 1997). The central service is a fault-tolerant system consisting of several computers plus links to information stored on servers housed at other locations. The communications options between remote computer and central service are varied, in part to serve the wide range of field conditions that exist at the data collection sites.

The central service plays several roles in supporting data collection, survey management and processing. One of the most critical functions with respect to developing Web-based survey tools is to store and serve the sample list and associated survey data. A complete listing of the data for each sample unit is always present on the server. Even if a sample unit has been "checked out" to a data gatherer's handheld computer for completing the CASI or for post-data collection editing, the most recent information on the sample unit is present in the database. This includes the survey data and codes that define the data collection site to which the sample unit is assigned, check-out status, data collection status, data gatherers and handheld computers that have been in contact with the sample unit, and dates of various actions, such as check-out and completion.

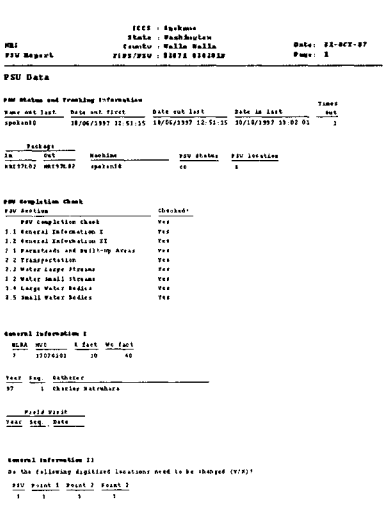
The initial CASIC system was designed solely for the purpose of supporting data collection and post-data collection processing. One area that quickly developed due to the existence of this system was serving information and interactive utilities via a Web browser. The main focus has been make materials accessible via desktop computers. However, an area of increasing interest is how to configure information for a handheld Web browser.

5 Web-based Survey Tools

The Web-based tools developed for the NRI consist of relatively static posted materials such as protocol descriptions, and dynamic products such as reports based on automatic daily database queries and interactive training. The tools support many aspects of the survey process, including preparation, data collection, survey management, and processing. Web-based dissemination tools are under development by NRCS. In the remainder of the paper, we describe tools developed for various phases of the survey. Except for the Web-based training, most of these tools are not publicly available on the Web.

5.1 Instructions and Training Materials

One of the earliest materials posted on the Web was the instruction manual for data collection. PDF (portable document format) files were created for each module, ensuring that all data gatherers had access to a document with identical formatting and content. Each document was labeled with version control information so that updates could be tracked. A user interface was developed to assist in viewing, printing and downloading the instruction manual or individual modules. This method of distributing instructions provided uniform electronic access to the most recent version of the instructions for all data gatherers, and



PSU Data

PSU Status and Tracking Information

Name sent last	Date sent first	Date out last	Date in last	Times out
spokane0	10/06/1997 12:51:35	10/06/1997 12:51:35	10/10/1997 10:00:01	1

Package In	Out	Machine	PSU Status	PSU Location
nat97L02	nat97M.02	spokane0	co	1

PSU Completion Check

PSU Section	Checked
PSU Completion Check	Yes
1.1 General Information I	Yes
1.2 General Information II	Yes
1.1 Perennials and Built-Up Areas	Yes
2.2 Transportation	Yes
2.2 Water Large Streams	Yes
1.2 Water Small Streams	Yes
1.4 Large Water Bodies	Yes
2.5 Small Water Bodies	Yes

General Information I

NLRA	MOS	% Sect	Wt fact
7	17074101	10	40

Year	Seq.	Gatherer
97	1	Charley Matchbox

Field Visit

Year	Seq.	Date

General Information II

Do the following digitized locations need to be changed (Y/N)?

PSU	Point 1	Point 2	Point 3
1	1	1	1

Figure 2. First page of a data view to assist in monitoring a data gatherer's work and to examine data as part of the editing process.

When the user requested data for a PSU, a program was run which created an approximately ten-page report in PDF format that showed the value for each variable in the survey for a specified PSU and its associated points. The supervisor could then either view or print the report depending on their needs.

Other utilities were developed to assist systems staff and data gatherers. For example, a session monitor for server activity was developed to assist systems staff in monitoring the central service's performance and to permit data gatherers to determine the current server load.

5.3 Monitoring Survey Progress

Web-based map and tabular displays of survey progress were created using status information that was derived from the central database. These reports included summaries of the number of completed PSUs and percentage completion for the entire nation, by state and by data collection unit (Figure 3 and 4). These initial reports led to development of more specific summaries for a variety of purposes. For example, a report was created summarizing the progress of data collection in relation to a planned data collection schedule for each ICCS; this was monitored by management to see which states were keeping up with production goals and which required additional attention to reach data collection goals. Data collection production reports by data gatherer (i.e., by

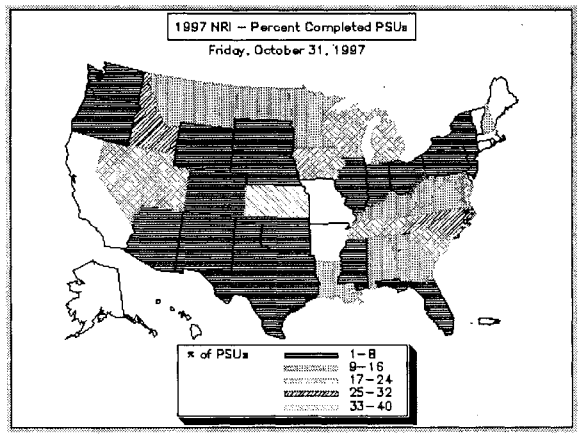


Figure 3. Map of data collection progress by state in the early months of data collection for the 1997 NRI.

National Resources Inventory
State by PSU Status Summary

Date:12-MAY-98
Page:1

Number of PSUs
(Row Percent)

State	IN-NOT STARTED	RETURNED UC	OUT-NOT RETURNED	RETURNED COMPLETE	Row Total
Alabama	1289 (21.3)	14 (0.2)	20 (0.3)	4729 (78.1)	6052
Alaska	42 (3.2)	1 (0.1)	10 (0.8)	1264 (96.0)	1317
Arizona	1223 (42.5)	10 (0.3)	24 (0.8)	1624 (56.4)	2881
Arkansas	1845 (30.0)	11 (0.2)	109 (1.8)	4178 (68.0)	6143
California	3011 (34.7)	14 (0.2)	186 (2.1)	5458 (63.0)	8669
Caribbean	1596 (66.2)		37 (1.5)	777 (32.2)	2410
Colorado	3578 (46.0)	332 (4.5)	28 (0.4)	3520 (47.2)	7458
Connecticut	231 (18.8)		35 (2.9)	962 (78.3)	1228
Delaware				498 (100.0)	498

Figure 4. Example of a tabular progress report showing the number and percentage of PSUs that had not been started, were in progress, needed additional assistance, or were complete.

handheld computer id) were also created for staff management purposes. Progress reports were generated automatically by running batch jobs at night to update the information used in displays and tables, or created on-the-fly in response to a user's request. The tabular displays were created as PDF documents and could be viewed or printed by the users. The survey progress reports as a whole had a very positive impact on the ability for a ICCS leader to manage their site and in meeting data collection deadlines.

enabled corrections to the instructions to be made in a simple and centrally-controlled fashion.

Training materials (lesson plans, Powerpoint slide shows, handouts, etc.) were also posted on the Web in preparation for or shortly after training occurred. There were four centralized "train the trainer" sessions to teach site managers and senior data gatherers how to collect NRI data. By posting materials to the Web, trainees could review materials prior to attending the training session and it was simple for them to download the same materials for use in training staff at their data collection sites.

This method of training can become problematic if satellite training sessions at data collection sites are not conducted in the same fashion as the centralized training. In addition, for rapidly deployed surveys, it is sometimes difficult to schedule centralized training sessions. To increase the consistency of training across data collection sites and across projects, interactive Web-based training was developed as a pilot project for the 1997 Special NRI Study (Shinn, 1998), which involved photo-interpretation of approximately 6000 PSUs on high resolution photography. The interactive utility borrows for two schools of thought in learning theory: constructivism and behaviorism. Behaviorism assumes that the teacher is the expert and that material needs to be presented in a prescribed sequence in order for students to learn. In contrast, constructivism is a student-centric theory, which centers on the notion that students must interact with and manipulate materials in order to fully learn a subject. In sample surveys, there is a need to provide materials in a prescribed order and to establish the notion of the trainer as the domain expert in order to instill the proper ethic for rigorously collecting data according to the defined protocols. On the other hand, photo-interpretation is a difficult method of data collection (as is conducting a complicated and/or lengthy interview), and interaction with the basic materials is essential for the gatherer to understand the data collection methods.

Prototype modules for completing the CASI and for photo-interpretation and measurement were developed and tested during the 1997 Special NRI Study field season (Figure 1). For the 1998 Special NRI Study, this pilot experience was used as the basis for creating a production-level interactive training tool for all PSU measurements (see www.statlab.iastate.edu/survey/NRIAI/TRAINING/SPECIAL98/Intro.html). The measurement protocols were integrated with instructions for completing the CASI, and the interactive utility was included as an integral part of the centralized data gatherer training.

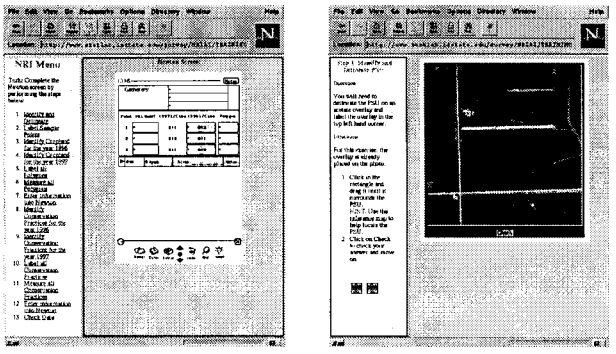


Figure 1. Interactive training modules for using the CASI and for photo-interpretation within the PSU.

Both the 1997 prototype and 1998 production-level interactive training module were developed using Java and standard Web programming tools (e.g., html). The system was designed so that the training utility would be run on the local computer to provide acceptable processing speeds in environments with low speed connections to the central server.

5.2 Data Collection Support

In addition to instructions and training materials, Web-based utilities were created to support data gathering. First, a Help Desk was established that included access to materials and announcements, as well as interactive tools. One of the most essential features of the Help Desk was that data gatherers could submit questions via e-mail to a team of technical staff with expertise in protocols and equipment usage. PDF documents with answers to technical questions were developed and posted to the Web so that a consistent set of answers was available to all data gatherers. Technical question/answer pages as well as a list of frequently and recently asked questions were developed so that they could be searched by users to determine whether an answer to a particular question already existed on the Help Desk site. This tool provided an increased level of consistency in providing technical support to data gatherers. An electronic chat room was also available for discussing issues related to the non-technical aspects of data collection, such as logistical strategies for staff management.

A second Web-based data collection tool was a "data view" to allow a supervisor to monitor a fraction of each data gatherer's work (Figure 2). The tool was developed to view PSU and point data without downloading the sample unit from the central server.

5.4 Post-Data Collection Processing

Because this is a longitudinal survey that is released as microdata, post-data collection editing is quite extensive. Warnings and error messages relating to consistency among variables are generated when PSU and point level edits are run, and additional areas for investigation are generated via diagnostic checks during statistical processing (e.g., imputation and weight calculation) and from summary table reviews. As a result, many waves of edits are sent to data gatherers to recheck and confirm or edit data. Tools developed to monitor the data collection process were extended to track the progress of edit checks sent back to data collection sites. Reports by data collection site and by individual PSUs were automatically generated on a daily basis or created in response to a user request.

As with data collection, Web-based systems tools were created to assist data editing staff. For example, a utility for altering the completion status of PSUs was developed to simplify requests to check out completed PSUs to update values in relation to central edit checks, and a utility was created to view data gatherer's ad hoc notes for a PSU when reviewing areas of concern generated during editing and processing of completed data.

6 Conclusions and Future Directions

Our work has shown that Web-based tools are easily integrated into CASIC systems. The tools may take the form of relatively static materials that are posted and occasionally updated, summaries generated from the survey database, or interactive utilities. If a centrally-accessible survey database is a component of the CASIC system, the variety of tools that can be developed is greatly expanded. In many survey operations, it is customary to provide interviewers (i.e., client computers) with their entire workload for the survey period. However, the timeliness of the information generated from a central database is optimized under the model in which the survey database is seen as the primary repository from which only a small number of sample units is checked out at one time by a single data gatherer.

The function of traditional survey materials is enhanced when coupled with the opportunities provided by the Web. The Web offers excellent accessibility to distributed sites. Because of this, it is much simpler to update materials and maintain consistency of information across sites and data gatherers than when using paper materials. The ability to automatically generate progress reports has greatly improved the efficiency of survey management and has provided a very important link between survey managers, funding organization representatives, and data gatherers in working together to achieve project goals.

Although we have focused on a large-scale application in natural resource surveys, it is possible to develop Web-based tools for many types and sizes of survey projects, depending on the type of staff and computing resources available. Static or occasionally updated materials can be posted to a Web page with standard computer-based office tools and staff that are facile with, for example, Windows software packages. A database manager is required to create a system to generate automatic progress reports, but standard database and visualization software can be used to create the maps and reports. The most difficult challenge is that of interactive training, which requires a staff member with reasonably advanced computer skills (e.g., Java) and knowledge of educational methods working closely with subject matter specialists.

Many of the Web-based tools developed represent natural extensions of existing paper-based materials traditionally used in the survey process. As more experience is gained in Web technologies, the form of these tools can be expected to change to take advantage of unique features provided by a Web-based interface. The interactive training is one step in this direction. Much work is needed to understand how one might provide these tools on smaller interfaces, such as a handheld computer environment. Finally, the NRCS is developing Web-based dissemination tools, not only for distributing the final 1997 NRI database and associated metadata, but to interactively generate analyses from the database using appropriate survey estimators.

Acknowledgments

This research was partially funded by cooperative agreement 68-3A75-7-90 between Iowa State University (ISU) and the USDA Natural Resources Conservation Service (NRCS). The authors wish to acknowledge several collaborators in developing these tools: Masoud Kazemi (ISU) and Herb Wilson (NRCS), who designed the monitoring reports; Gail Shinn and Pete Boysen (ISU), Bob Dayton and Tom O'Connor (NRCS), who developed the interactive training utilities; and the NRCS Resources Inventory Support Branch, which created the Help Desk pages for technical support and downloading the survey protocols and training presentations.

References

Nusser, S. M., D. M. Thompson, and G. S. DeLozier. 1996. Using personal digital assistants to collect survey data. 1996 Proceedings of the Section on Survey Research Methods, American Statistical Association, p. 780-785.

Nusser, S. M. and J. J. Goebel. 1997. The National Resources Inventory: a long-term multi-resource monitoring programme. Environmental and Ecological Statistics 4:181-204.

Nusser, S. M. and D.M. Thompson. 1997. Networked hand-held CASIC. Proceedings of *Symposium 97: New Directions in Surveys and Censuses*, Statistics Canada (forthcoming).

Shinn, G. R. 1997. The Development of the 1997 National Resources Inventory (NRI) Web-based Training Tutorial Project. Creative component, Iowa State University, Ames, IA.