

Development of an integrated CARI Interactive Data Access System for the US Census Bureau

Mai Nguyen, M. Rita Thissen, B.Christopher Siege, Sandhya Bikmal, RTI International

Abstract

In preparation for the American Community Survey Content Test, a team from the US Census Bureau and RTI International has been collaborating in the development of the CARI Interactive Data Access System, an integrated web-based system for reviewing of audio and image recordings which are made with Blaise 4.8.2 during field and telephone interviewing.

The use of computer audio-recorded interviewing (CARI) techniques is expanding and can be expected to grow rapidly with the release of CARI capabilities in Blaise software. Capture of these sound and image files is only the beginning of the process, however; to be useful, those recordings must be reviewed and analyzed in light of the study goals. CARI can have many values: confirmation of data quality and authenticity, examination of the effectiveness of question wording, study of interviewer and respondent behavior, managing performance of the interviewing staff, and other purposes. In this paper we discuss the design and implementation of the integrated CARI Interactive Data Access System with emphasis on data integration to bring audio and image recordings created from Blaise 4.8.2 during data collection into the system for quality review, authenticity evaluation and behavior coding of interview administration, respondent reactions and questionnaire design

Introduction

Blaise software offers many built-in capabilities for survey data collection, and with the release of version 4.8.2, the language now has the capability to record audio and capture snapshots of screen images during the course of an interview, storing them on the computer's local drive along with the response data. Whether used for in-person interviewing on laptops or computer-assisted telephone interviewing with voice-capture, the CARI-equipped Blaise instrument appears and behaves just as it would without audio or image capture.

By programming a survey questionnaire to take advantage of the audio and image recordings, researchers can obtain insight into the interviewing environment in detail through later review of the audio recordings and their matching screenshots. Many uses have been found for computer audio-recorded interviewing (CARI), ranging from quality control <Thissen et al, 2010> <Mitchell et al, 2008> <Thissen and Nguyen, 2008> <Basson, 2005> to coaching <Biemer et al, 2000> <Watt, 2009><Thissen et al, 2008> to data capture <Edwards et al, 2010> .

Over the past several years, the US Census Bureau has evaluated the use of CARI, with the intent of using it for production survey data collection after establishing its feasibility <Biemer et al, 2000><Wrenn-Yorker and Thissen, 2005>, acceptance by field staff <Arceneux, 2007> and cost-effectiveness <Biemer et al, 2002>. In 2008, a prototype CARI System was created at the Census Bureau <Thissen et al, 2008>. Stemming from that effort and other methodological and technical studies, a decision was made to develop a full-fledged system for use on the upcoming American Community Survey Content Test (ACS Content Test).

The ACS Content Test is programmed in Blaise 4.8.2, collecting audio and screen images as described in a separate paper at this conference, *Implementing Computer-Audio Recorded Interviewing (CARI) Using Blaise 4.8.2*, William E. Dyer, Jr. and Malcolm Robert Wallace, U.S. Census Bureau. To effectively use CARI recordings, the organization requires a secure system for receipt of the incoming audio and image

files, an efficient interface for playback and coding their contents, and especially for a large organization such as the Census Bureau, a facile mechanism for managing operations which will ultimately need to handle many thousands of cases passing through it on a flow basis.

The CARI Interactive Data Access System (CARI System) was collaboratively developed by RTI International and the Census Bureau for this purpose. The system was designed to reside centrally at the Suitland, MD, headquarters of the Census Bureau while being accessible to Census Bureau telephone centers in Hagerstown, MD, Tucson, AZ, and Jeffersonville, IN, as well as to the twelve regional offices across the country. To enable such broad access without distributed storage, the system provides a web based interface, requiring only a standard browser and network access credentials for usage.

The ACS Content Test is relying on CARI recordings from the Blaise instrument to aid with decisions on questionnaire design through behavior coding of the captured audio and image files. Later surveys expect to use the system for quality assurance, management and coaching of field staff, and potentially other purposes. In the pages that follow the CARI System's design, functionality, usage and implementation details are presented.

Functionality

The CARI System presents a web based interface currently consisting of 4 major components. Plans exist to expand the system to enable additional functionality. Two of the components provide direct access to the CARI recordings from Blaise, while two others support daily operations.

- **Behavior Coding Component (BCC).** The BCC is designed to allow researchers at the US Census Bureau to capture the interactions between interviewer and respondent by listening to the CARI audio files and viewing images captured during data collection. Coding interviewers' behaviors as part of the survey interaction not only allows research managers to monitor whether interviewers consistently adhere to the required (standardized) interviewing practices, but it also alerts them about possible problems with the questionnaire. For example, deviations from the interviewing script may reflect weaknesses in the questionnaire design or skip logic. On the other hand, the respondents' behaviors may also reveal if respondents adequately understand and respond to a particular question. The request for explanation, providing inadequate response, or other non-scripted discussions may indicate a weakness in the comprehensibility of the respondent task or survey question.
- **Quality Assurance Component (QAC).** Quality assurance activities ensure that the data collected in a survey have as little error as possible. Survey error can arise from numerous sources, but generally fall into two main categories: authenticity problems and errors introduced by individual differences in presentation of the questionnaire by survey interviewers.

The first error source, relating to authenticity, arises when interviewers do not collect data directly from a respondent. This may occur in an innocent way, as in the case when interviewers feel they already know the respondent's answer and do not wish to burden them by asking the question, or it may arise as a result of malfeasance, in which the interviewer does not collect data at all or fabricates it. In any of these cases, the data cannot be considered authentic.

The second source of error, resulting from differences in presentation of questions to the respondent, can be seen as a lack of precision. Most survey data collection is predicated on the notion of consistent delivery of the questionnaire to minimize individual differences. Though computerization forces some degree of uniformity, the interviewer may still cause variation through failure to follow protocol (such as non-verbatim reading), emotional loading (such as

through tone of voice), biasing the response by suggestion or leading comments, or other violations of best interviewing practices. These faults weaken the reliability and validity of the data set as a whole, though individual data points may be correct.

The QAC is designed to allow monitors to evaluate the quality of collected data by listening to the audio files and viewing images captured during data collection.

- **Automatic Assessment Tool for User Feedback (AAT).** The AAT module allows the Census Bureau staff to collect user feedback to evaluate ease of use and effectiveness of the CARI system from the user's perspective. RTI and Census Bureau staff will use the feedback to identify future improvements to the CARI system.
- **eTraining.** The eTraining module provides integrated, online training materials to introduce new coders to the system and to train them on how to manage their workload and to conduct the behavior coding and/or quality assurance coding properly.

Coding Workflow

The behavior and quality assurance coding processes are managed by Census Bureau staff at various divisions following specific workflows associated with each component. In general, the overall division of labor among the Census Bureau divisions is:

- Site supervisors and staff at the 3 telephone sites will primarily use the website to conduct the behavior and quality assurance coding activities.
- The survey configuration, case assignment to sites, workload monitor, and data extraction for analysis are done by research staff at the HQ.
- Website administration and its infrastructure are maintained and supported by Census Bureau staff.

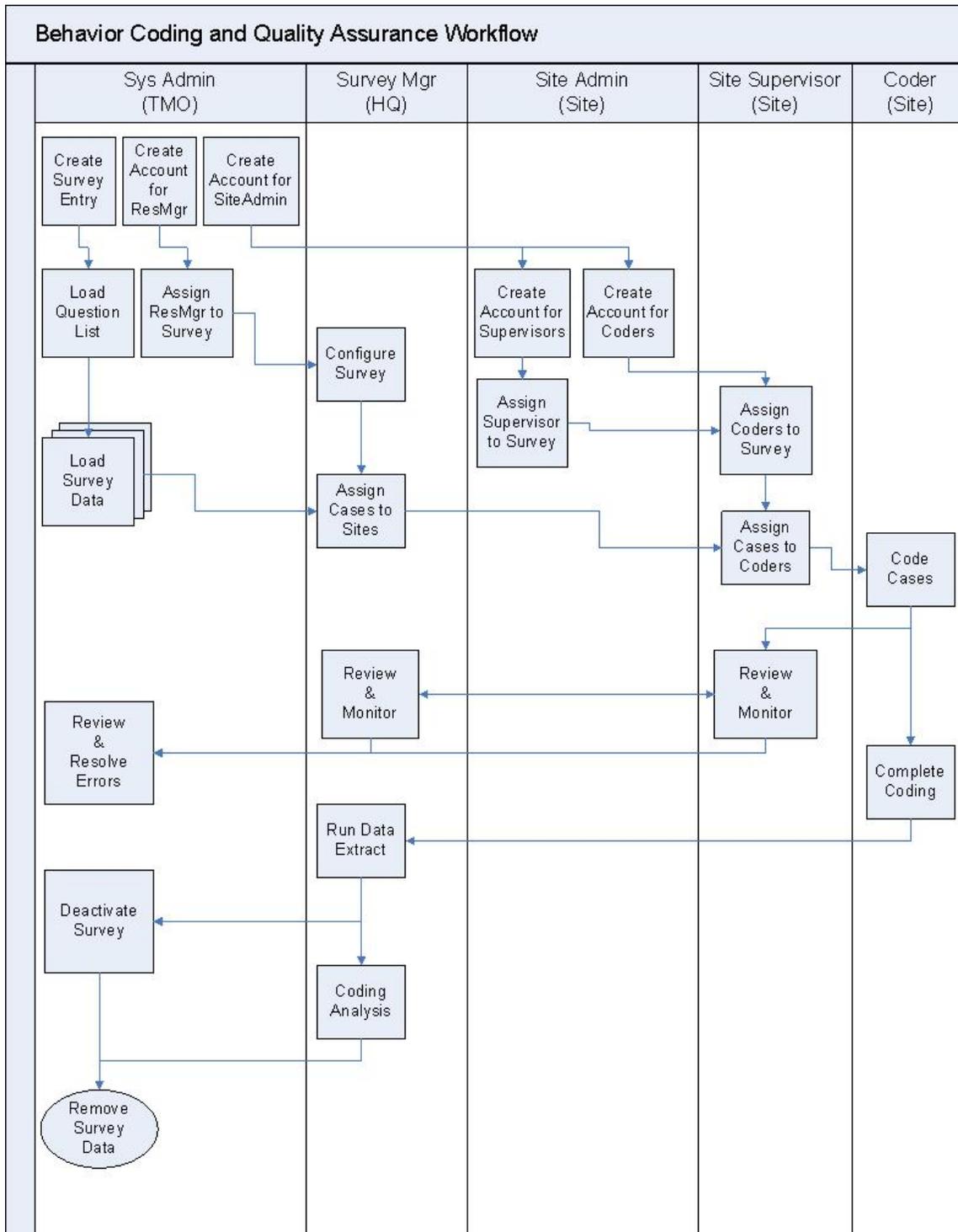
The CARI System is designed to support this division of labor through the use of user roles as shown in **Figure 1**. The coding workflow includes all tasks from the creation of a new survey entry through coding completion and finally survey data removal, in chronological order, identifying the responsible party at each step. The major tasks in this work flow are:

- Create new survey entry and user accounts
- Configure and assign staff to the survey
- Load data into the system
- Assign cases to sites and coders
- Conduct coding
- Monitor coding progress
- Data extraction and close-out

Each user role has different responsibilities within the website, providing a clear division of labor from setup through closeout. The original setup for any survey takes several steps, after which operations become more routine. To begin with, survey entries are created in the system by the System Administrator. Subsequently, a list of the questionnaire items which were recorded is uploaded through an interface provided as part of the system. Once a survey entry is ready, the case data can be loaded into the system, but it cannot be coded yet. The System Administrator at this time may create accounts for users, such as a Survey Manager and Site Administrator, giving them access to the newly created survey entry. Responsibility then shifts to the Survey Manager, who configures the coding interface with survey-specific header items, coding categories, coding groups and code definitions. The Research Manager also designates groups of cases for any sites at which coders will work, after which the Site Administrators can

assign those cases to specific Coders, who listen to audio, view screen images, select appropriate codes and write notes with their observations. The majority of work is done by the Coders, who may review thousands of cases for a single survey, and Supervisors, who oversee the work of the Coders. When all coding is complete, the Research Manager and System Administrator again take responsibility, determining when to close out the survey entry. These steps are very similar for both behavior coding and for quality assurance coding, providing a workflow which is fully managed through the CARI website interface.

Figure 1. CARI System Workflow, with Activities Identified by User Role



The Behavior Coding page, together with the Window Media Player for playing audio file and the captured images, is shown in **Figure 2** below. Through this page, the coder can access the audio recording by clicking on the Audio link and view the image by clicking on the Image link. The page also provides

the capability for flagging the case if further review is needed. Further, there is provision for additional notes on the case or for each utterance to be captured by the coders.

Figure 2. Behavior Coding Page, with Window Media Player and Captured Image

CARI
Interactive Data Access
User: coder001 Component: Behavior Coding

Role: Coder
Survey: RTI Test 50

[Log Out](#)

[Work Load](#)

[-] CNTR1037

- ✓ Sample Quest...
- ✓ Sample Quest...
- ✓ Sample Quest...
- Sample Quest...
- Sample Quest...

[+] CNTR1038

[+] CNTR1039

[+] CNTR1040

<< Previous
Next >>
Save
Reset
Mark Case Complete

Survey ID:	RTI_0050	Survey Name:	RTI Test 50
Control Num:	CNTR1037	Question:	Sample Question
Coding Status:	In Progress	Date/Time Last Saved:	08/31/2010 01:48:51 PM
Assigned Coder ID:	coder001	Assigned Coder Role:	Coder
Audio Link:		Audio Date/Time:	04/28/2008 12:47:39 PM
Image Link:		Image Date/Time:	04/28/2008 12:47:39 PM
Data Collection Mode:	0	Interviewer ID:	intv001

Reason For Not Coding:

Categories	Respondent Behavior	Language
Interviewer	R-AA Adequate Answer <input checked="" type="checkbox"/>	L-E English <input checked="" type="checkbox"/>
Respondent	R-CL Clarification <input type="checkbox"/>	L-M Mixture of English and Spanish <input type="checkbox"/>
Final	R-DK Don't Know <input type="checkbox"/>	L-PI Pronunciation problems for Intv <input type="checkbox"/>
	R-IA Inadequate Answer <input type="checkbox"/>	L-PR Pronunciation problems for Resp <input type="checkbox"/>
	R-IU Inaudible/Uncodable <input type="checkbox"/>	L-S Spanish <input type="checkbox"/>
	R-R Refuse <input type="checkbox"/>	
	R-UA Uncertain Answer <input type="checkbox"/>	

Codes: R-AA L-E

Notes: Lorem ipsum dolor sit
adipiscing elit. Donec
eget elementum dapibus

CCC_GetFile 00:02

Add Coding Entry

Category	1	2	3
Interviewer	I-ES	I-V+	B-I
Respondent	R-AA	L-E	

Case Notes:

<< Previous
Next >>

Disclosure Prohibited - Title 13 U.S.C.

Close Image Window

Forms Answer Navigate Options Help

We'll begin by selecting a sample of your residents. Because this is a statistic, provide a list of the current residents as of midnight yesterday. The sampling p initials of each resident. Each resident should be listed on a separate line.

Then, I will briefly review the list with you. Finally after I enter the total number c residents we will include in the sample.

PRESS *1* AND ENTER TO CONTINUE.

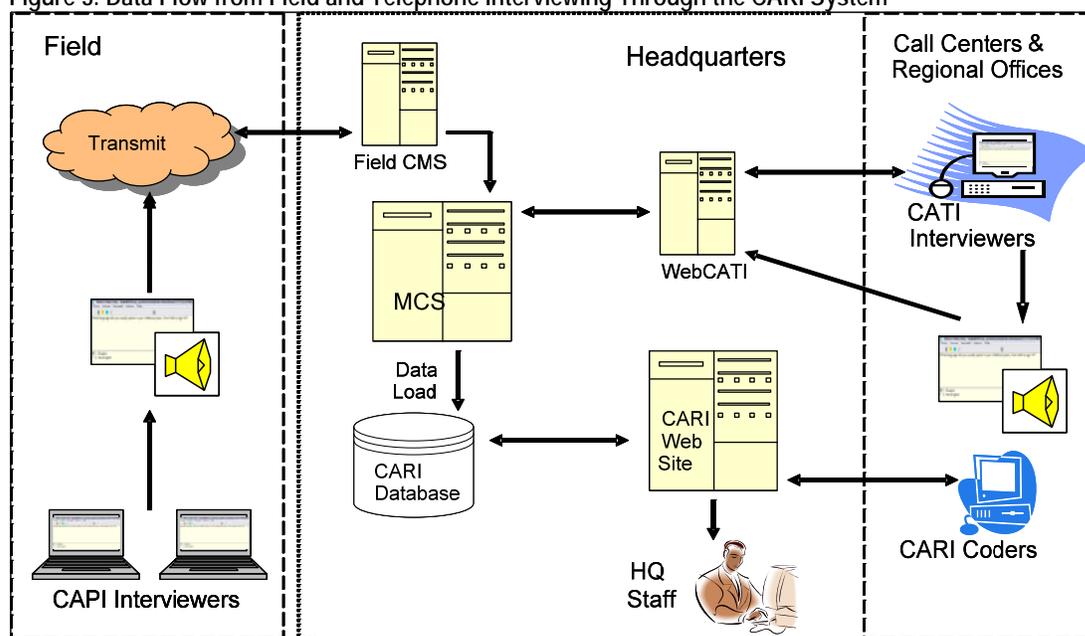
C1 CONTINUE

Data Flow

The purpose of the CARI System is to make Blaise CARI files available for study by staff who are located around the country, potentially quite distant from the interviewing sites or even from the Census Bureau's campus.

Response data files, including CARI files, originate at the point of interviewing, such as on the laptop of a field interviewer or the desktop of a telephone interviewer. Regardless of where the data start, the files are transmitted to a centrally located master control system (MCS) for loading into the CARI System. A simplified diagram of the overall data flow is shown in **Figure 3**. CARI coders may be located at any of the Census Bureau's locations, depending on the needs of the survey and the type of analysis for which the CARI files are being coded. Ultimately, the coded results become available to headquarters (HQ) staff, both as summary statistics and as exported datasets.

Figure 3. Data Flow from Field and Telephone Interviewing Through the CARI System



The CARI System utilizes an Oracle database as its data repository, which users of the CARI System access through the CARI website. CARI audio files and screen images, selected interview data and paradata flow into the database from the MCS through a pre-defined programmatic data-loading interface. The database also accumulates administrative data such as user accounts, roles, survey configuration parameters, code definitions and other operational information.

Once loaded, case information and CARI recordings become available to coding staff through the website. Codes are assigned, comments written and operational flags are set by the coding and supervisory staff through the web pages, which validate and store this data alongside the original information. CARI System reports and data review pages provide details and summaries to interactive users, and an extraction function allows authorized research managers to export datasets to secure locations for further analysis. Upon completion of the work, when there is no further need for storage, system tools enable removal of the unnecessary data.

System Design Overview

The CARI System has been developed using the Microsoft ASP.NET Framework 3.5 with an Oracle database. The Visual Basic.NET programming language is used for server-side development tasks in conjunction with the JavaScript language for handling client-side activities. AJAX technology is used in highly interactive pages to provide better user experience.

The Microsoft ASP.NET Framework provides a modern platform for developing enterprise-scale web applications. It provides the following development benefits:

- **Improve productivity** by using the form-based, drag-and-drop development model familiar to many developers. Microsoft Visual Studio 2008, being one of the best development tools, provides a further boost to productivity.
- **Leverage large infrastructure and prebuilt class libraries** available via the Microsoft .NET 3.5 framework, including ADO.NET for data access, server-side web UI controls, AJAX programming support, etc.
- **Support for enterprise relational database management system (RDBMS)**, including Microsoft SQL Server, Oracle database and IBM DB2.
- **Ease of deployment**, especially to Windows servers running the IIS web server.

The CARI System is hosted on the Census Bureau's Windows server using the Microsoft IIS web server. It is accessible within the Census Bureau secured intranet network. It has achieved the security certification and accreditation specified by NIST SP 800-37. The CARI System website is designed to be viewed with the Microsoft Internet Explorer (IE) web browser version 6.x or later.

In addition to network security, the CARI System also employs a number of security measures to control user access and protect its data:

- SSL protocol implementation.
- Integrated LDAP user authentication using the Census Bureau's eDirectory.
- Role-based authorization to control user access to both web pages and data within the system.
- Protections against SQL injection and cross site scripting attacks using best practices.
- Validation of all data input
- Use of parameterized queries and stored procedures for all database access
- Encryption of configuration data

The CARI System controls data access using two types of roles, system roles and survey-specific roles. System roles include:

- **System Administrator.** Members of this role will be able to access all web pages and view and modify all data for all surveys.
- **Security Administrator.** Members of this role will be able to access select web pages for security auditing purpose only. They will not be able to see any data within the system.
- **Site Administrator.** Members of this role will have access to all web pages, but can only view and modify user accounts for their site (user accounts) and cases assigned to their site.

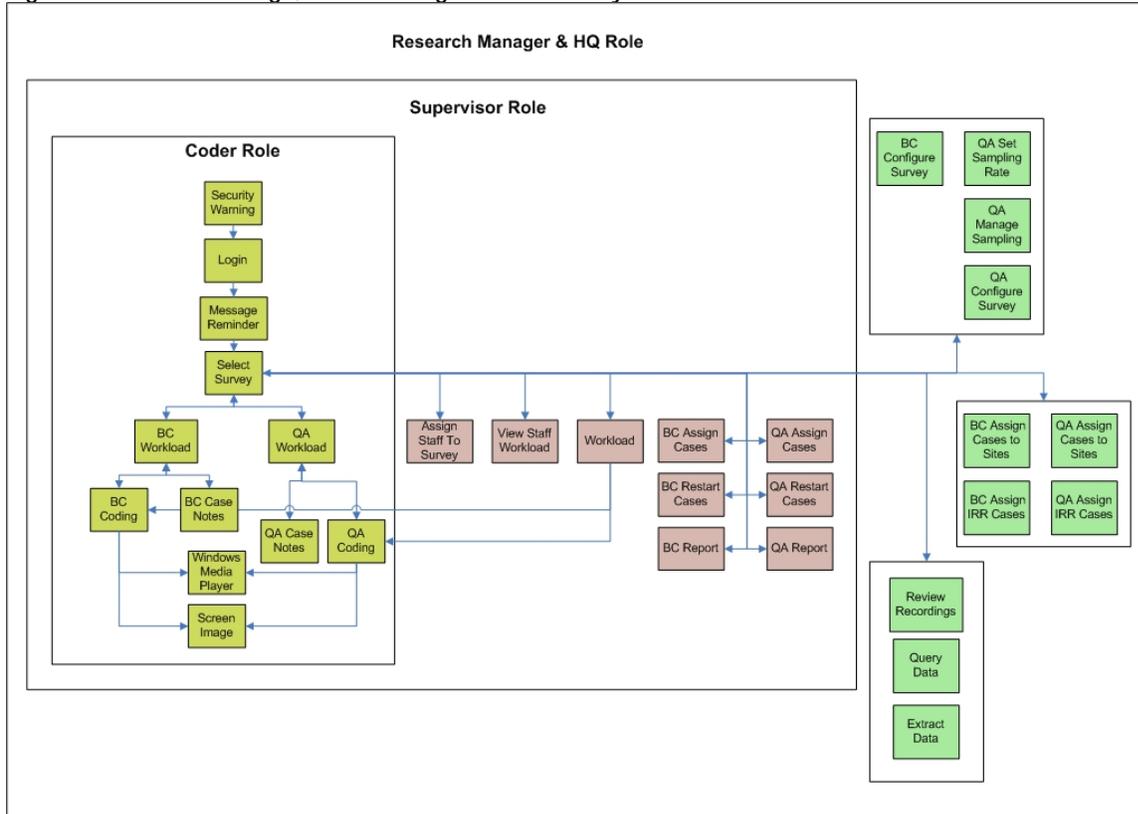
Survey-specific roles include:

- **Coder.** Members of this role can only access coding-related web pages and work on cases assigned to them. They can see neither cases from other users nor cases from surveys they're not assigned to.
- **Supervisor.** Members of this role can access coding-related web pages, workload management pages and reports, but they can only view and modify cases assigned to their site.

- **Research manager.** Members of this role not only can access all pages as a supervisor can but also can view and modify cases across sites for any survey assigned to them. They are able to access pages to configure these surveys and set up the QA sampling rate parameters.
- **HQ.** Members of this role have the same access level as research managers with the exception that they cannot modify any data.

This role-based design can be seen in **Figure 4** where each colored box represents a web page accessible by the specified role.

Figure 4 – User Role Design, with Web Pages Accessible by Each Role



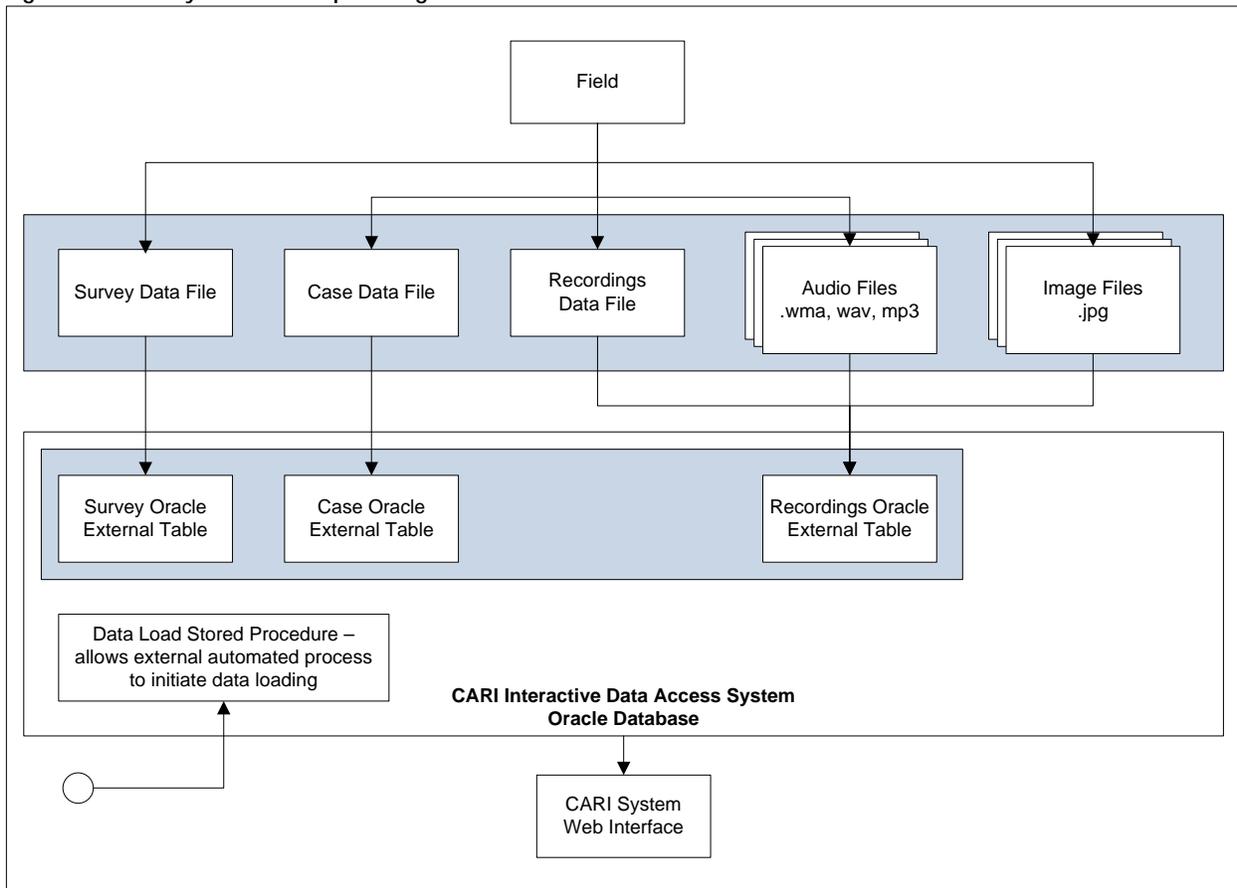
System Integration

The CARI System's primary point of integration is with the Census Bureau's master control system (MCS) through a data loading process. This process transfers selected paradata along with audio and images files from the interview case records and is capable of quickly loading thousands of recordings into the CARI System database. Within the CARI System web interface an initial survey configuration is defined and the instrument questions are imported. Once complete the CARI System can support an automated batch process of loading the survey case data either as a single complete set or on a continual flow basis.

Integration between the systems is facilitated through the use of Oracle external tables. These external tables allow flat files to be exposed within the Oracle database and provide a convenient way to process external data for importing. They provide an abstracted interface between the systems and allow details of the exporting process from the MCS to be decoupled from the importing process of the CARI System.

The definition of the flat files provides an interface contract between the systems as illustrated in **Figure 5**.

Figure 5 – CARI System Data Import Diagram



Once the data load process is complete the user can then access interview cases within the web based interface for either Behavior or Quality Assurance Coding. Within these primary areas of functionality, the user can access case and question level attributes while reviewing both the Blaise screen capture and audio recordings.

After case and question level review has been completed within the CARI System it is possible to perform a number of defined data extracts. Associated with most of the data extracts is a SAS program to allow for external analysis.

Conclusion

The new Blaise functionality of audio and image recording opens a wide range of potential opportunities for survey research and operations. As mentioned above, the technology enables behavior coding, quality management and coaching activities. Additional possibilities present themselves:

- Conducting “cognitive interviews” through silent observation of production interviewing, by use of CARI files.
- Data collection, for open-ended questions or for semi-structured (conversational) interviewing
- Enhanced training for interviewing staff, allowing them to practice interviewing and review their own performance

With further technological advances, speech analytics and transcription may become feasible for post-processing the recorded audio, and pen-based computers might be able to record signatures as part of the screen image collected for consent questions. These and other activities will make good use of Blaise's newly introduced capabilities.

References

Arceneaux, T. "Evaluating the Computer Audio-Recorded Interviewing (CARI) Household Wellness Study (HWS) Field Test." 2007 Proceedings of the American Statistical Association. Statistical Computing Section [CD-ROM]. Alexandria, VA: American Statistical Association, 2007. 2811-2818

Basson, Danna. 2005. "The Effects of Digital Recording of Telephone Interviews on Survey Data Quality." Pp. 3778-85 in the American Statistical Association 2005 Proceedings of the Section on Survey Research Methods. Washington, DC: American Statistical Association

Biemer, P.P., D. Hergert, J. Morton and W. G. Willis (2000), "The Feasibility of Monitoring Field Interview Performance Using Computer Audio Recorded Interviewing (CARI)", Proceedings of the American Statistical Association's Section on Survey Research Methods, pp. 1068-1073

Biemer, P.P., F. Mierzwa, M.R. Thissen (2002), "Comparison of the Steady-State Costs for CARI vs. Reinterview Approaches for Interview Verification", RTI Technical Report to the US Census Bureau

Edwards, Brad, W. Hicks and Michelle Carlson (2010, June). Old Dogs New Tricks: CARI and CASI Innovations to Reduce Measurement Error and Nonresponse, Presented at the International Total Survey Error Workshop, Stowe, VT.

Mitchell, S.B., K.M. Fahrney, M.M. Strobl, M.R. Thissen, M.T. Nguyen, B.S. Bibb, and W.I. Stephenson (2008). "Using Computer Audio-Recorded Interviewing to Assess Interviewer Coding Error," Presented at the 63rd Conference of the American Association for Public Opinion Research (AAPOR). New Orleans, LA USA

Thissen, M.R., Fisher, C., Barber, L., Sattaluri, S. (2008). "Computer Audio-Recorded Interviewing (CARI), A Tool for Monitoring Field Interviewers and Improving Field Data Collection", Proceedings of the International Methodology Symposium 2008, Statistics Canada, Gatineau, Canada

Thissen, M.R., & Nguyen, M.T. (2008, June). Using Audio Recording to Detect Data Collection Error. Presented at The 2nd International Total Survey Error Workshop, Research Triangle Park, NC.

Thissen, M.R., Thorpe, S., Peytcheva, E., Barnes, C., Park, H., & Fisher, C. (2010, June). Improving Data Quality Through Audio and Image Recording. Poster presented at International Total Survey Error Workshop, Stowe, VT

Wrenn-Yorker, C and Thissen, M.R. (2005), "Computer Audio Recorded Interviewing (CARI) Technology", Presented at the Federal Computer Assisted Survey Information Collection (FedCASIC) Conference.

Watt, Barbara (2009, May), "Accountability for Data Quality Excellence", Presented at the International Field Directors and Technologies Conference, Delray Beach, FL

Acknowledgements

We would like to thank the U.S. Census Bureau for funding the development of the CARI System. We would also like to thank the Census Bureau staff for their contribution and collaborative effort in this development, as well as a number of RTI staff including Carl Fisher, Neelima Kunta, Brett Anderson, Erica Saleska and Emily McFarlane Geisen. The views expressed in this paper are those of the authors and not necessarily those of the U.S. Census Bureau.

Contacts

Mai Nguyen
RTI International
919-541-8757
mnguyen@rti.org

B.Christopher Siege
RTI International
919-485-5605
csiege@rti.org

M. Rita Thissen
RTI International
919-485-7728
rthissen@rti.org

Sandhya Bikmal
RTI International
919-386-0502
sbikmal@rti.org