Innovative Web Based Documentation System Designed for the National Survey on Drug Use and Health

Nanthini Ganapathi, Susan K. Myers, Inga B. Allred (RTI International)

nanthini@rti.org, smyers@rti.org, irb@rti.org
RTI International, 3040 Cornwallis Rd, PO Box 12194, Research Triangle Park, NC 27709-2194

Abstract
Data documentation is equally important to the success of a study as the data themselves. It is imperative that data collected and created within studies be thoroughly documented so that they can be analyzed. Developing high quality metadata is a tough challenge for any information management community. When these metadata span several study years and involves a large volume of data, the need arises for a centralized location allowing users to add, modify and track variables.

The National Survey on Drug Use and Health (NSDUH) is a nationwide survey conducted annually by RTI International for the Substance Abuse and Mental Health Services Administration (SAMHSA). It collects interview data from approximately 67,500 respondents in all 50 states and the District of Columbia. The NSDUH is the primary source of information on the prevalence, patterns, and consequences of alcohol, tobacco, and illegal drug use and abuse in the general U.S. civilian non institutionalized population, age 12 and older. There are currently multiple data analyses active at different stages on the NSDUH, and each one may have as many as 3,000 variables to document.

RTI International has developed web based Variable Tracking System (VTS) software for variable documentation and reporting for the NSDUH project. The goal was to create software that would keep current documentation available to analysts at all times and automate the creation of electronic and paper codebooks. The VTS is an efficient and user friendly system allowing users to enter, edit and search for information about variables. The VTS also allows users to identify the individuals responsible for the variable and its documentation, so that the responsible person can be contacted for questions. In addition, this system serves as the source of all metadata that are assimilated into deliverable documentation for the client. This paper focuses on the key features of VTS and discusses its future direction.

Background
RTI has conducted the NSDUH annually since 1988. Each study year has three data files and codebook deliverables. They may be thought of sequentially as Raw followed by Analytic and finally a Public Use File. Each codebook contains the metadata of the data file. Elements of metadata may include any or all of the following:

- Section heading
- Variable Name
- Comment or note about the variable
- SAS label
- Frequency levels (distinct, range or combination of each)
- Frequency level definitions
- Associated footnote

An excerpt from the NSDUH code book is given in Figure 1.

```
                        NON-CORE DEMOGRAPHICS                          PAGE:    603
LABEL   LEN   DESCRIPTION                                                 FREQ     %
-----   ---   -----------                                                 ----     -
  (QD17)      NOTE:  The school enrollment variable SCHENRL is
              analogous to the variable ENROLED from prior years.
              SCHENRL takes into account information from
              follow-up probes if respondents originally
              reported that they were not enrolled in school.
              Based on these probes, respondents were coded as
              yes (i.e., enrolled) if they were on break from
              school and intended to return to school once their
              break was over.  SCHENRL was assigned a code of
              11 if the respondent reported currently being
              enrolled but there was uncertainty about the
              respondent's school enrollment status.

              The next questions are about school.  Are you now
              attending or are you currently enrolled in school?
              By "school," we mean an elementary school, a junior
              high or middle school, a high school, or a college
              or university.  Please include home schooling as well.

 SCHENRL  2   NOW ENROLLED IN ANY SCHOOL
               1 = Yes.....................................   26263  48.56
               2 = No......................................   27765  51.34
               5 = Yes LOGICALLY ASSIGNED (LFSCHWH2=601)...      12   0.02
              11 = Yes (SCHDSKIP = 30)....................       8   0.01
              94 = DON'T KNOW.............................       4   0.01
              97 = REFUSED................................       4   0.01
              98 = BLANK (NO ANSWER)......................      23   0.04
```

Figure 1. Excerpt from Codebook

With approximately three thousand variables for each study year and codebook type, the volume of metadata is a challenge to manage. Most of the analytic and imputed variables are created by analysts and keeping track of them manually is difficult. Prior to the use of the VTS, metadata were stored in a database not accessible to the analysts. Since there was no user interface, analysts did not have ready access to the current documentation and had to rely on outdated codebooks for information. Since the documentation originates with the analysts, the descriptive information was passed along with the variable itself to a programmer who updated the database. Obviously this was a labor intensive, cumbersome and error prone process. The VTS has allowed us to bypass this step and improve data quality and efficiency. End users include analysts who merely view the available metadata, staff that deliver variables and associated documentation to the master database, and programmers that extract relevant metadata used in the deliverable codebooks.

Technical Aspects

Figure 2 below shows the technical implementation details of the VTS. As mentioned earlier, the VTS is a client-server type web-based system which users can access through any web browser. The user interface was developed using a combination of ColdFusion, HTML, and client side JavaScript. The VTS web site is hosted on RTI's corporate web server cluster. The back end for the VTS is a relational database hosted on RTI's corporate SQL Server. When users request codebooks, the web interface records the request in the SQL Server database. An automatic job scheduler, running on a windows-based desktop class application server, examines these requests every sixty seconds, and initiates SAS jobs to create appropriate codebooks for all pending requests. The completed codebooks are emailed to the requester, who will typically receive them within minutes of their original request. At the current time, this system resides behind RTI's firewall and is inaccessible to anyone outside of the RTI network. For additional security, VTS is also password protected within RTI.

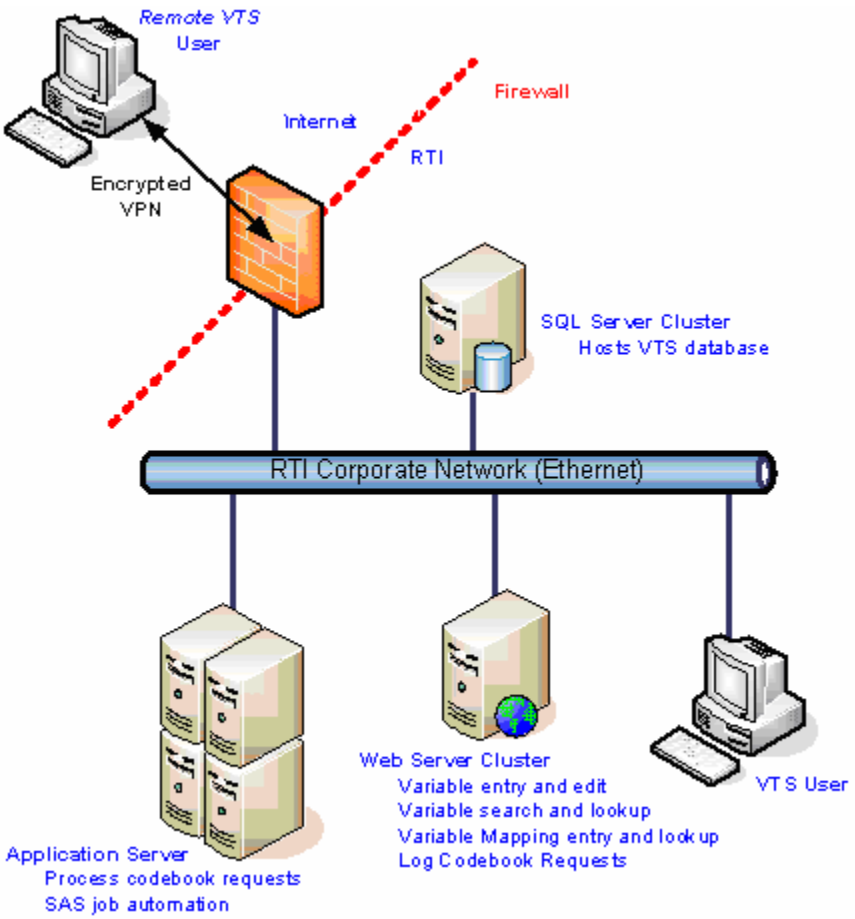


Figure 2. Technical infrastructure of the VTS

Key Features

The VTS is accessed from a main web page illustrated in Figure 3 below. The features described in this section are accessed by links on this page. The key features of VTS are

- Variable Delivery History
- Documentation Entry / Search
- Automated Code Book Creation
- Interrelationship or Mapping of Variables

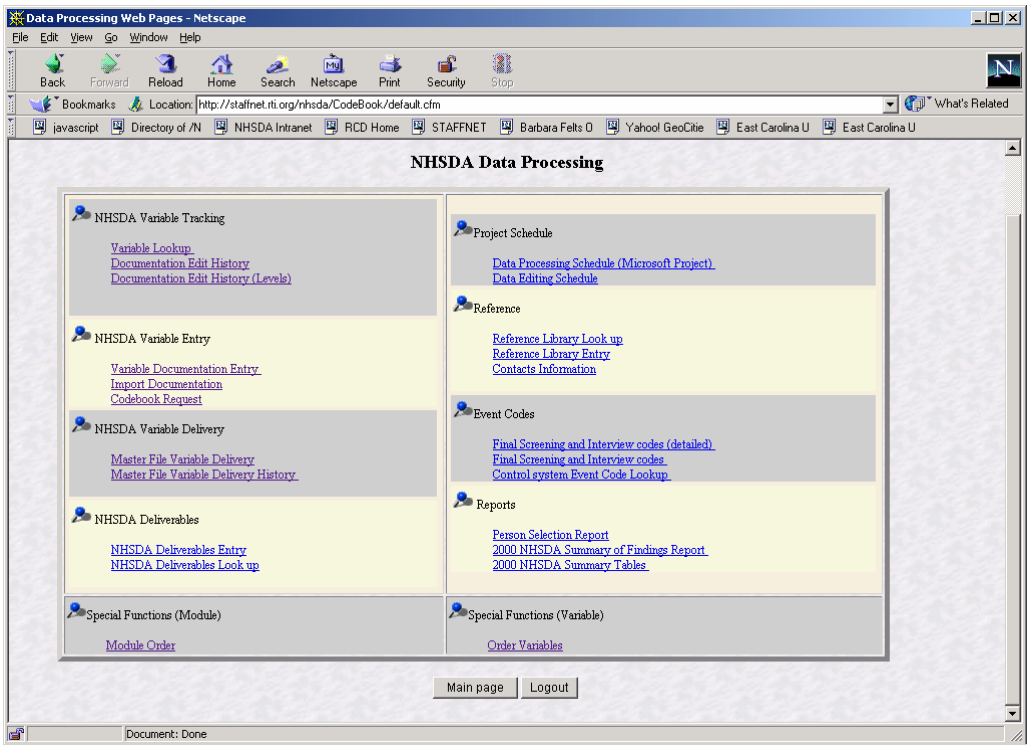


Figure 3. VTS Main Page

Variable Delivery History

Variables are delivered through the VTS and automatic notification emails are sent to a team distribution list. This serves as a notification for the master database manager to merge the associated SAS dataset with the master SAS dataset. Sometimes variables are "redelivered" to the master database when errors are detected or modifications are requested. These are flagged as "redelivered" so the history of the variable may be tracked. Automatic notification emails to a specific distribution list indicate these redelivered variables clearly, so that the other related variables can be recreated.

The VTS has a filter that may be used to see the life history of a variable. As illustrated in Figure 4, variable deliveries may be filtered by many fields including delivery dates. This is particularly useful when codebooks are redelivered. With this feature, it is easy to determine what has changed on the data file since a specific date. Our example shows the original delivery on 02/18/04 and subsequent redelivery on 02/24/04. There is a hyperlink in the 'Notes' field that brings up any important comments the analyst may have regarding the (re)delivery.

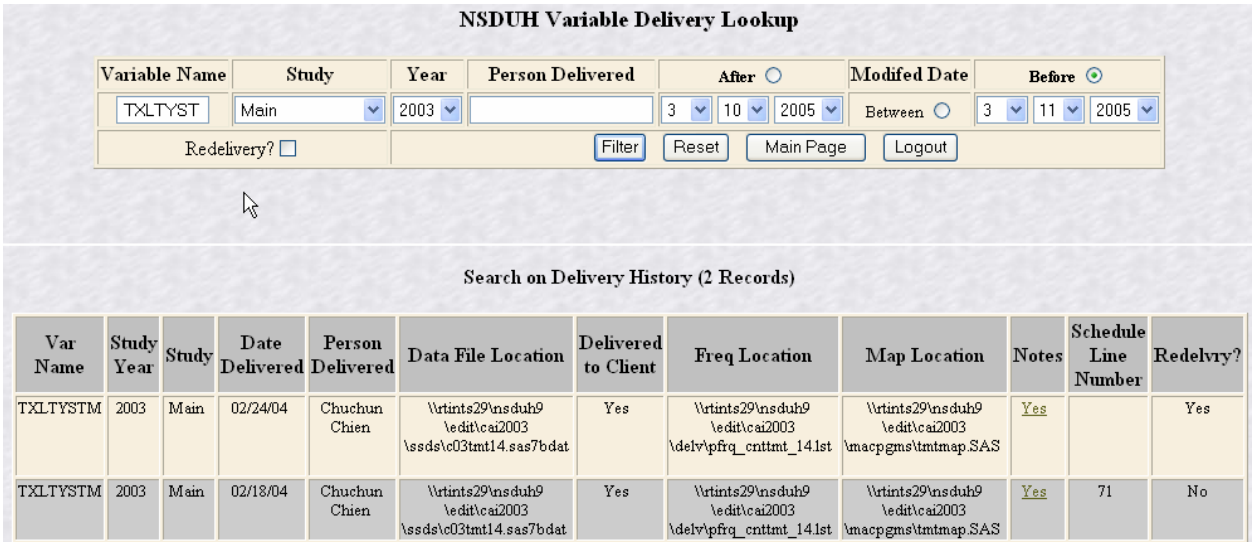


Figure 4. Delivery Log

Documentation Entry / Search

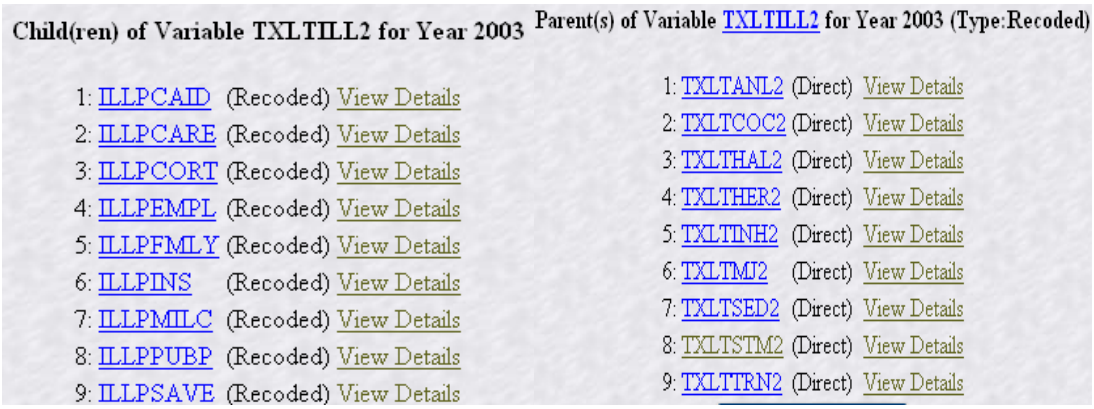
Descriptive data can be entered or edited in the system as variables are delivered. This is the place for metadata entry mentioned earlier as components of a codebook: variable name, comments, SAS label, frequency level definitions, footnotes and section headings. Moreover, the system provides a powerful searching capability to search for any specific variable or keyword in the documentation. The system maintains a comprehensive change history for each variable, which can be viewed within the system.

Automated Codebook Creation

An important feature of VTS is the ability to create codebooks automatically. Since metadata are displayed in a data entry format on the web, it is important for analysts to see how it appears in a codebook format. To ensure that all pertinent information is present and correct, the user requests a codebook online by choosing a set of variables. This request is processed on the server with a SAS program. The resulting codebook is emailed to the requestor promptly. This process has greatly improved the quality of deliverable codebooks. In the old system, whole codebooks, sometimes 800 pages long, were sent to staff for review before delivering to the client. Breaking down the review to smaller sections prior to the full review has greatly improved the quality and overall efficiency of the process.

Interrelationship or Mapping of Variables

Relationships among variables can be complex. As mentioned previously, some variables are derived from others. The natural flow is from Raw to Edited to Analysis to Imputed. Any number of preceding variables may be used in the creation of the next generation variable. Analysts use this information in several ways. An important use of this data is to assess the impact of redelivered variables. Many times as variables are modified, it is necessary to know what other variables may be affected. For this reason, a means of entering parent, children and sibling variables is available.



Benefits

The VTS has proven to be a valuable tool for our project staff including the analysts and the programmers. It has many benefits which have already been mentioned. In addition, the following features of the system contribute to its success as a data quality tool.

- Multiple Users
- Quality Assurance
- Central Location available to everyone for use at any time
- Data Tracking
- Keyword searching
- Copy features
- Quicker response time to Client's requests

The web based VTS allows multiple users to enter and view data simultaneously from any location with internet access and permissions to access RTI's internal network. Users do not have to install any specific software in their computers to access the system. Through the use of record level locking, staff are able to enter data without the risk of their updates being overwritten by another user.

The VTS allows analysts to enter their own documentation, which greatly reduces the risk of data entry errors. It also provides them with the ability to review and modify documentation as needed.

The VTS serves as a repository for all documentation. It is secured with Usernames and Passwords. A person with access can retrieve all documentation associated with a variable, including labels, level documentation, question text, and frequencies at any time. The Keyword Search feature allows the user to search thousands of lines of documentation for a keyword in seconds.

The VTS allows staff to see the "history" of the variable; if and when it was delivered to the Master file, along with all of its documentation and documentation history. This aids in notifying responsible parties when a variable and/or its documentation is in question.

In the NSDUH project, most variables exist from one study year to the next. For these variables, though the data change, the metadata often remain the same. In this case, there is a 'Copy' feature that allows the analyst to pull all documentation from one year to the next for selected variables. For multi-year studies like NSDUH this is a necessary feature so that data entry is done only for new variables.

Finally, the VTS makes it possible for staff to respond much more quickly to ad hoc codebook requests.

Future Enhancements

We hope to eventually enhance our system by providing a centralized, web based system for tracking questionnaire specifications. This will help to track questionnaire changes electronically in a common database instead of several emails and documents. This would enable the initial loading of question (comment) text as well as variable names into the VTS, which would reduce the risk of data entry errors by eliminating the need for data entry by multiple staff. Currently, instrument specifications are sent to programmers in a MS Word document. To initialize documentation within the VTS, one uses the Word document as a guide and pastes question text and variable names into the system. If the original specifications were captured with a database driven means, it would require a simple upload of metadata into the VTS.

The VTS is an evolving system, with new features added as users identify needs. The programming staff hopes to continue this trend as new staff work with it and communicate new ideas for the system.