Paper 161-31

Data Profiling Using Base SAS® Software: A Quick Approach to Understanding Your Data

Susan J. Nowlin, National Institute for Occupational Safety and Health,
Cincinnati, OH

ABSTRACT

"Data Profiling is the use of analytical techniques about data for the purpose of developing a thorough knowledge of its content, structure and quality" (www.bitpipe.com). While this terminology is most often associated with Data Warehousing and high-level business intelligence efforts, these techniques are valuable tools for the everyday data manager or data analyst. SAS® Version 9 software offers various avenues for performing data profiling such as, SAS/ETL and SAS Data Quality Solution. These tools however, may not be available for some SAS users, may require additional training, and may be overkill if an understanding of the content of a file is all that is needed; that is, no data cleansing or other transformations are required. This paper discusses an application using only base SAS software which provides basic statistics, frequencies, ranges, outlier, and structural information for each variable in a table. The result of the application is a condensed report detailing the information about the content of a data file. The application was written using the Windows environment and can be run from the SAS Display Manager. For those who have SAS/IntrNet® software, a front end is also available to provide a user friendly interface. Current enhancements under development include running the application from SAS Enterprise Guide® as a stored process.

NIOSH disclaimer: The findings and conclusions in this paper have not been formally disseminated by the National Institute for Occupational Safety and Health and should not be construed to represent any agency determination or policy.

Introduction

As part of performing epidemiological research studies at the National Institute for Occupational Safety and Health (NIOSH), it is often required that the study data managers pull together data from multiple sources in order to create the databases necessary for final analysis. These sources of data may be files in various formats received from work sites, other government agencies, data coded from hardcopy, or may be legacy files stored in NIOSH archives from previous studies. It is the data manager's responsibility to obtain an understanding of all data sources available for a study and to work with the study team to determine which pieces of which files need to be merged, combined or otherwise manipulated to create the analysis files. Since many files are received from sources outside of NIOSH, the data quality can not be relied on. In the case of legacy files, from within or outside NIOSH, data dictionary information and/or data business rules are often sketchy or non-existent.

Using individual SAS PROCs and SAS statements to analyze the data content is useful but not always efficient. Running simple PROC FREQs on an entire table to look for outliers produces many pages of output, especially for variables which have many unique values. Other information that is needed (field name, field type, field length, field description, number of unique values, range of values, frequency of individual field values and the number of missing values) requires running additional SAS code, in effect producing even more pages of output. Trying to organize all these individual pages of output is inefficient and cumbersome to review.

A Quick Data Profiling Solution

The application presented in this paper was originally written in 1995 using base SAS statements, PROC SQL, PROC CONTENTS, and the Macro facility to produce a condensed report of some basic statistics about the data in each field of a given table. The program was written to aid NIOSH data managers in obtaining a quick report about the electronic data received from Human Resource, Industrial Hygiene and Health Physics departments of work sites being studied. The goal of the output was to provide the information in a format as concise and condensed as possible.

The solution involved developing an application that would run the necessary SAS statements in order to gather the information about the data, then output the information in a customized report. Options were added to the application using macro variables to select a table for processing, select the variables to process, restrict the number of frequency lines printed, select sort order of variables, and to produce a sample listing of the file if desired. If the application is run in SAS Version 8 or earlier, the input file must be a SAS dataset, imported to SAS, or read using one of the SAS/ACCESS® tools. If the application is run in SAS Version 9, SAS will automatically use the appropriate libname engine for SAS datasets, Microsoft Access tables or Excel files so no additional tools or processing is required for these file types.

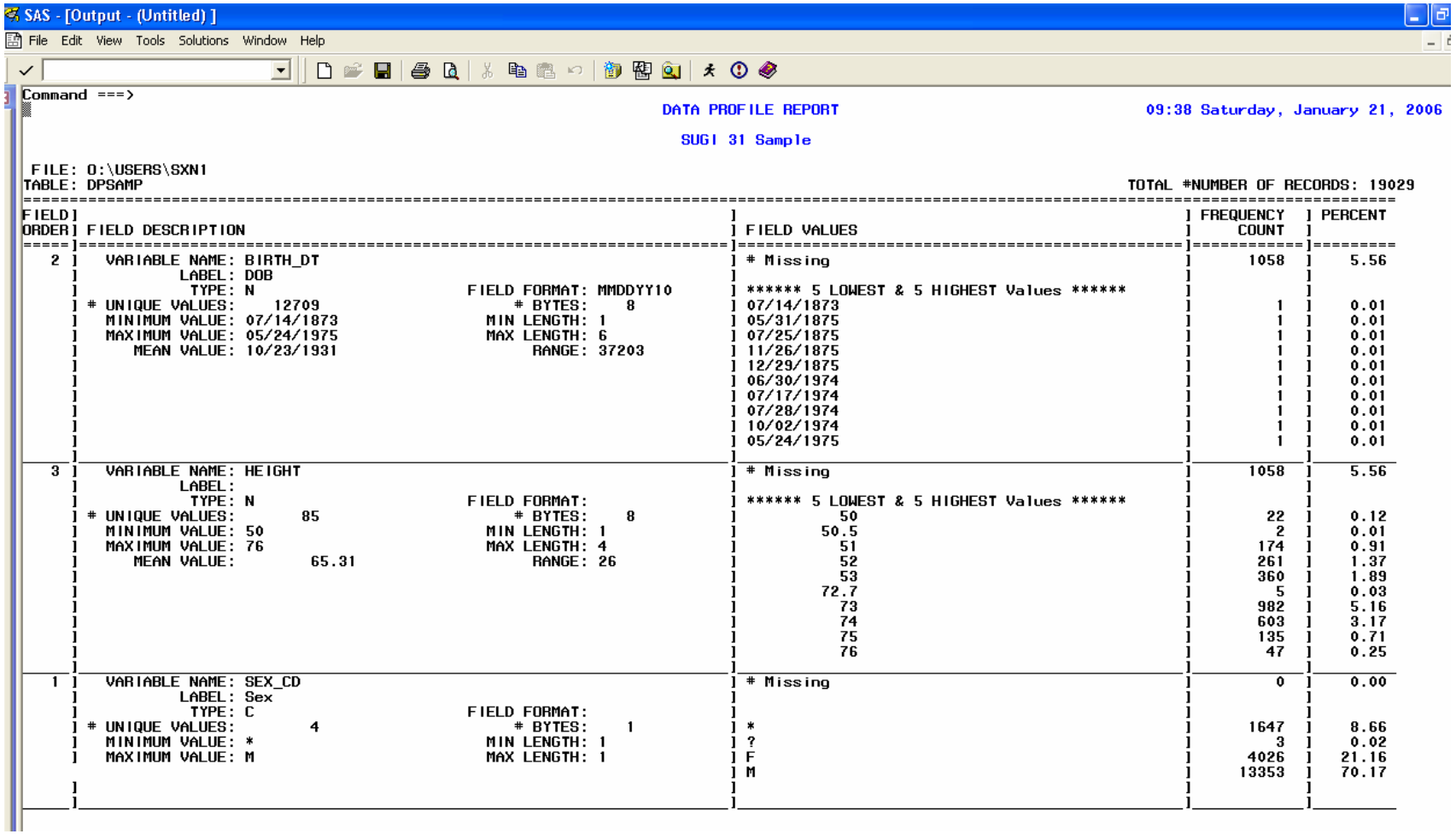
Since 1995, improvements have been made including the addition of a user friendly interface which is run using SAS/IntrNet. The output report for this interactive interface is produced using HTML and is displayed in a web browser. An example of each version of the program is demonstrated in this paper. Current enhancements under development include the ability to run the application as a stored process from SAS Enterprise Guide.

Display Manager Interface Example

From the SAS Display Manager's program editor window, the user enters information for the Data Profiling program options by modifying the contents of the Macro variables in the following front-end program:

```
*--------------+-------------------------------------------------------------------------+
| PRODUCT      | NIOSH DATA PROFILING APPLICATION
| FILENAME     | DATA_PROFILE_DM.SAS
| DESCRIPTION  | This program produces a report on the content of a given data file.
| VERSION      | 2.0
| LANGUAGE     | SAS/BASE
+--------------+-------------------------------------------------------------------------+
| PURPOSE      | To provide detailed information about the content and structure of each field
|              | in a data table in order to assess data quality.
+--------------+-------------------------------------------------------------------------+
| PARAMETER    | FILEPATH:  Full path name where the data file to be analyzed is located.
|              | TABLE:     Name of the table (file) to be analyzed.
|              | TASKNO:    (Optional) IT Task Tracking Number.
|              | TITLE:     (Optional) User Report Title.
|              | VARLIST:   (Optional) Enter a list of variables to include or exclude from
|              |            the report.
|              | VARSTAT:   (Optional) Enter 'KEEP' to Include the variables in the VARLIST,
|              |                       Enter 'DROP' to Exclude the variables in the VARLIST.
|              | MAXPRINT:  Establishes the maximum number of frequency lines to print per
|              |            variable.
|              | SORTBY:    Enter 'VARNUM' to sort output by variable position in the table,
|              |            Enter 'VAR' to sort output by variable name.
|              | SAMPSIZE:  Enter the number of records to print in a sample print out of
|              |            of data table, or 0 (zero) to supress the sample print.
+--------------+-------------------------------------------------------------------------+
*_____________________________________________________________;
*ENTER THE FULL PATH WHERE THE DATA FILE IS LOCATED;
%LET FILEPATH=O:\USERS\SXN1;

*_____________________________________________________________;
*ENTER THE NAME OF THE FILE (TABLE) TO ANALYZE;
%LET TABLE=DPSAMP;

*_____________________________________________________________;
*ENTER THE IT TASK TRACKING NUMBER (OPTIONAL);
%LET TASKNO=;

*_____________________________________________________________;
*ENTER A TITLE FOR THE REPORT (OPTIONAL);
%LET TITLE=SUGI 31 Sample;

*_____________________________________________________________;
*ENTER VARIABLES TO INCLUDE/EXCLUDE IN THE REPORT (OPTIONAL);
%LET VARLIST=BIRTH_DT HEIGHT SEX_CD;

*_____________________________________________________________;
*ENTER 'KEEP' TO INCLUDE ABOVE VARIABLES OR
*      'DROP' TO EXCLUDE ABOVE VARIABLES          (OPTIONAL);
%LET VARSTAT=KEEP;

*_____________________________________________________________;
*ENTER 'VARNUM' TO SORT FREQ REPORT BY VARIABLE POSITION ON THE FILE OR
*      'VAR'    TO SORT BY VARIABLE NAME;
%LET SORTBY=VAR;

*_____________________________________________________________;
*ENTER MAXIMUM NUMBER OF DETAIL FREQUENCY LINES TO PRINT PER
* VARIABLE.;
%LET MAXPRINT=40;

*_____________________________________________________________;
*ENTER NUMBER OF RECORDS TO PRINT IN THE SAMPLE PRINTOUT OF
*THE DATASET (ENTER 0 TO SUPPRESS);
%LET SAMPSIZE=50;

*_____________________________________________________________;
%INCLUDE "L:\INFO_TECH\DATA_PROFILING_APPLICATION\INCL_DATA_PROFILE_ANALYSIS.SAS";
RUN;
```

After modifying the macro variables, the user submits the program from the SAS Display Manager. The following report is displayed in the Output Window of Display Manager:



SAS - [Output - (Untitled)]

File Edit View Tools Solutions Window Help

Command ===>

DATA PROFILE REPORT 09:38 Saturday, January 21, 2006

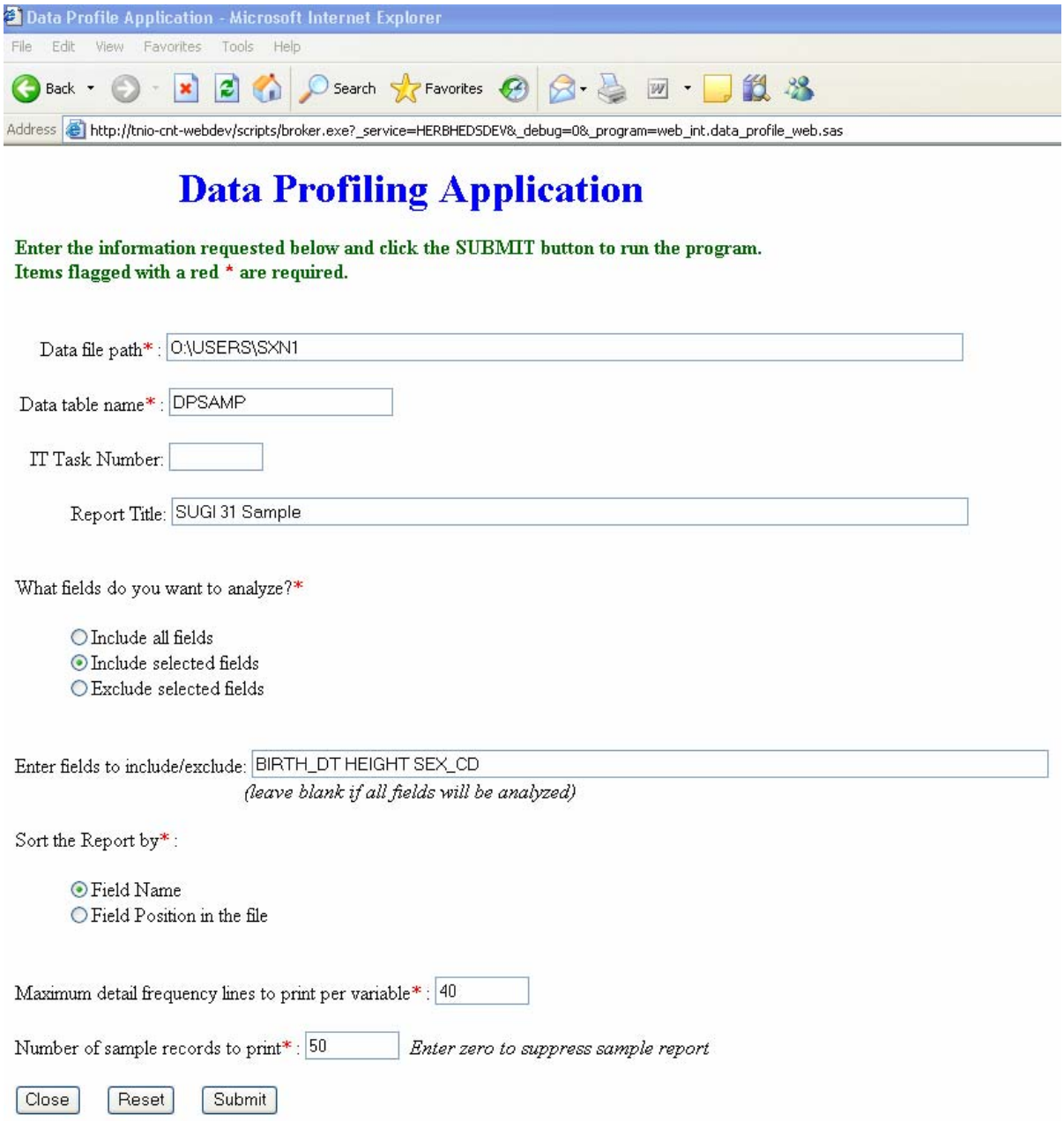
SUGI 31 Sample

FILE: O:\USERS\SXN1
TABLE: DPSAMP TOTAL #NUMBER OF RECORDS: 19029

FIELD ORDER	FIELD DESCRIPTION		FIELD VALUES	FREQUENCY COUNT	PERCENT
2	VARIABLE NAME: BIRTH_DT		# Missing	1058	5.56
	LABEL: DOB				
	TYPE: N	FIELD FORMAT: MMDDYY10	****** 5 LOWEST & 5 HIGHEST Values ******		
	# UNIQUE VALUES: 12709	# BYTES: 8	07/14/1873	1	0.01
	MINIMUM VALUE: 07/14/1873	MIN LENGTH: 1	05/31/1875	1	0.01
	MAXIMUM VALUE: 05/24/1975	MAX LENGTH: 6	07/25/1875	1	0.01
	MEAN VALUE: 10/23/1931	RANGE: 37203	11/26/1875	1	0.01
			12/29/1875	1	0.01
			06/30/1974	1	0.01
			07/17/1974	1	0.01
			07/28/1974	1	0.01
			10/02/1974	1	0.01
			05/24/1975	1	0.01
3	VARIABLE NAME: HEIGHT		# Missing	1058	5.56
	LABEL:				
	TYPE: N	FIELD FORMAT:	****** 5 LOWEST & 5 HIGHEST Values ******		
	# UNIQUE VALUES: 85	# BYTES: 8	50	22	0.12
	MINIMUM VALUE: 50	MIN LENGTH: 1	50.5	2	0.01
	MAXIMUM VALUE: 76	MAX LENGTH: 4	51	174	0.91
	MEAN VALUE: 65.31	RANGE: 26	52	261	1.37
			53	360	1.89
			72.7	5	0.03
			73	982	5.16
			74	603	3.17
			75	135	0.71
			76	47	0.25
1	VARIABLE NAME: SEX_CD		# Missing	0	0.00
	LABEL: Sex				
	TYPE: C	FIELD FORMAT:			
	# UNIQUE VALUES: 4	# BYTES: 1	*	1647	8.66
	MINIMUM VALUE: *	MIN LENGTH: 1	?	3	0.02
	MAXIMUM VALUE: M	MAX LENGTH: 1	F	4026	21.16
			M	13353	70.17

SAS/IntrNet® Interface Example

The user enters the options for the Data Profiling program through the following interface:



After all options are selected and the Submit button 

is clicked, the following report will be displayed in the user's web browser:

Data Profile Report

DATASET:DPSAMP

TOTAL NUMBER OF RECORDS: 19029

VAR #	VARIABLE DESCRIPTION	VARIABLE VALUES	FREQUENCY COUNT	PERCENT
2	Variable Name: BIRTH_DT	# Missing	1058	5.56
	Label: DOB	** 5 LOWEST & 5 HIGHEST Values **		
	Length: 8	07/14/1873	1	0.01
	Type: N	05/31/1875	1	0.01
	# Unique Values: 12709	07/25/1875	1	0.01
	Minimum Value: 07/14/1873	11/26/1875	1	0.01
	Maximum Value: 05/24/1975	12/29/1875	1	0.01
		06/30/1974	1	0.01
		07/17/1974	1	0.01
		07/28/1974	1	0.01
		10/02/1974	1	0.01
		05/24/1975	1	0.01
3	Variable Name: HEIGHT	# Missing	1058	5.56
	Label:	** 5 LOWEST & 5 HIGHEST Values **		
	Length: 8	50	22	0.12
	Type: N	50.5	2	0.01
	# Unique Values: 85	51	174	0.91
	Minimum Value: 50	52	261	1.37
	Maximum Value: 76	53	360	1.89
		72.7	5	0.03
		73	982	5.16
		74	603	3.17
		75	135	0.71
		76	47	0.25
1	Variable Name: SEX_CD	# Missing	0	0.00
	Label: Sex	*	1647	8.66
	Length: 1	?	3	0.02
	Type: C	F	4026	21.16
	# Unique Values: 4	M	13353	70.17
	Minimum Value: *			
	Maximum Value: M			

Click here to view the sample data

SAS Data Profiling Tools Comparison

The NIOSH Data Profiling application provides the following information about each field in a table:

- Data Structure (field name, type, length, label, format)
- Frequency count of each unique value (aids in outlier detection)
- Incomplete / missing data
- Number of unique values
- Minimum and Maximum values
- Minimum and Maximum length of field content
- Mean value and Range of values for numeric fields
- Five lowest and five highest values (if the frequency output restriction is applied).

The SAS Data Quality Solution and ETL software packages are much more robust, providing the following additional data profiling analysis:

- Detailed metadata validation (primary key candidate, null count, etc.)
- Pattern analysis (used to determine if the data values in a field are in the expected format)
- Expanded statistical analysis for numeric data fields (median, mode, std. dev, etc.)
- More advanced frequency count and outlier information
- Rule Validation (canned techniques including range checking, look-up validation or specific formulas, and the ability to store and validate against an organization's unique business rules)
- Relationship Discovery (information about logical connections between pieces of data)

Conclusion

Obtaining a clear understanding of the content of a data source is invaluable to a data manager regardless if the task at hand is simply to evaluate the quality of a given data file or to develop a large scale - metadata driven - data warehouse using ETL processes. Using the appropriate tool for the job is what is often difficult to decipher. The power tool is *always* more exciting to use but is it necessary and/or cost effective? (When the power is needed and available, however, SAS Data Quality tools are highly recommended).

The benefits of data profiling are numerous:

- Aids in quickly determining the quality of a data source.
- Facilitates the discovery of nuances, discrepancies, inaccuracies, outliers and gaps in data.
- Doesn't allow you to *assume* quality. Garbage can be discovered *before* a data source is used for a project.
- Provides a quick insight to multiple data sources when several are being considered for use on a project.
- Can provide some documentation previously unavailable or unknown on legacy data files. Particularly valuable when no one is around any longer to ask questions.
- Greatly facilitates data cleansing and transformation efforts…..and should always be the first step on such projects.

The benefits of using the NIOSH Data Profiling application are:

- It is easy to use, no special training is necessary.
- It provides many data profiling techniques for those who do not have the more powerful SAS Data Quality tools.
- It could be used as a first step in a larger scale data quality analysis project even if the more powerful SAS Data Quality tools are ultimately planned to be used.

References

http://www.bitpipe.com/rlist/term/Data-Profiling.html
http://www.sas.com/technologies/dw/dataquality/
http://www.dataflux.com/Data-Management/Data-Profiling/index.asp
"SAS Data Quality – A Technology Overview", Eric Hunley, SAS, SUGI 29

Acknowledgements

The HTML and SAS/IntrNet portions of this application were developed by Jun Ju, Constella Group. Other enhancements made during the evolution of this application were contributed by Kathy Waters of NIOSH and Yanmei Li of Constella Group.
Thank you to BJ Haussler, Steve Ahrenholz, and Doug Daniels of NIOSH and Glen McCann of UGS for reviewing the content of this document.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.

Contact Information

Susan J. Nowlin
IT Specialist
National Institute for Occupational Safety and Health
4676 Columbia Pkwy, MS-R4
Cincinnati, OH, 45226
(513) 841-4467
snowlin@cdc.gov