**Paper A14-2009**

# The Super Genius Guide to Generating Dummy Data

Brian Varney, COMSYS, Portage, MI

## ABSTRACT

A common necessity in development and programming is having representative data to program and develop against. Situations that can hamper this could be that the data has not been collected yet, the data is too sensitive to share or just lack of resources to prepare and provide the data from the source data tables. This paper is intended to provide methods to generate representative data whether one has the project data or only metadata in such a way that sensitive data is not revealed

## INTRODUCTION

Often times, development or programming is delayed because of the lack of representative data. This paper is intended to briefly present a few different types of and possible methods for generating sample data to use for development.

The audience for this paper should have at least an introductory knowledge of SAS/Base and SAS Macro in order to follow along.

At this time, I consider this paper a work in progress as I am sure I will receive good ideas and feedback. I plan on using the ideas and feedback to further mature this paper to be more complete and useful for other SAS users. I apologize in advance to any data that is offended by being referred to as “dummy data”.

## ATTRIBUTES OF GOOD DUMMY DATA

This section intends to discuss attributes that can make your dummy data more useful for development.

### DATA STRUCTURES MATCH

This is really a necessity. If the field names and attributes do not match, development will really be hampered.

### DATA VALUES ARE VALID (OR PURPOSEFULLY, INVALID)

Development is much more meaningful if the data values are representative. For example, the field gender should be mostly M/F or Male/Female. If it makes sense, a few outliers, invalid values and missing values should be mixed in. If the data field is a comment field, the varying value lengths should be used in the data generation.

### RELATIONSHIPS BETWEEN KEYS ARE REPRESENTATIVE

Primary keys and foreign key relationships between tables are important to represent in an accurate manner to have a legitimate development environment.

### SECURITY OF SENSITIVE DATA IS UPHELD

If the data contains items such as social security numbers, names, addresses, etc., it is not recommended to try and use it for development or testing in a shared environment. Sensitive data needs to be closely monitored and should not be copied from the production environment.

### THE DATA IS EASILY AVAILABLE TO SAS ENVIRONMENTS

There are certain data sets that come with SAS by default. Some of them depend on the modules that are licensed. Good development data should be easily available to the development, testing and production environments. There should not be data that is too sensitive to be made available to those who need it.

## POSSIBLE SOURCES FOR DUMMY DATA

Since there are many situations and requirements around data, different types of dummy data are more appropriate in different situations.

### SASHELP DATA SETS

As all SAS users know, the SASHELP library is always there. If one is providing some sample code to show SAS functionality from a training perspective, what better data set to use than one such as SASHELP.CLASS? Another option is to use some of the views against the dictionary tables such as SASHELP.VTABLE a.k.a. DICTIONARY.TABLES.

### GENERATING WITH THE DATA STEP

The data step is perfectly suited for generating data due to the flexibility of the functions available, the natural looping mechanism and the ability to output several data sets at once.

### GENERATING FROM METADATA

By leveraging the dictionary tables containing the metadata about the production data, one can generate dummy data of the exact same structure.

### GENERATING EMPTY TABLES WITH PROC SQL

The following PROC SQL code can be used to create empty tables with the same structure as the source data tables.

```
proc sql noprint;
   create table <target_table> like <source_table>;
quit;
```

### GENERATEDATA WEB SITE

This is a thin client tool to help generate a single table of dummy data.

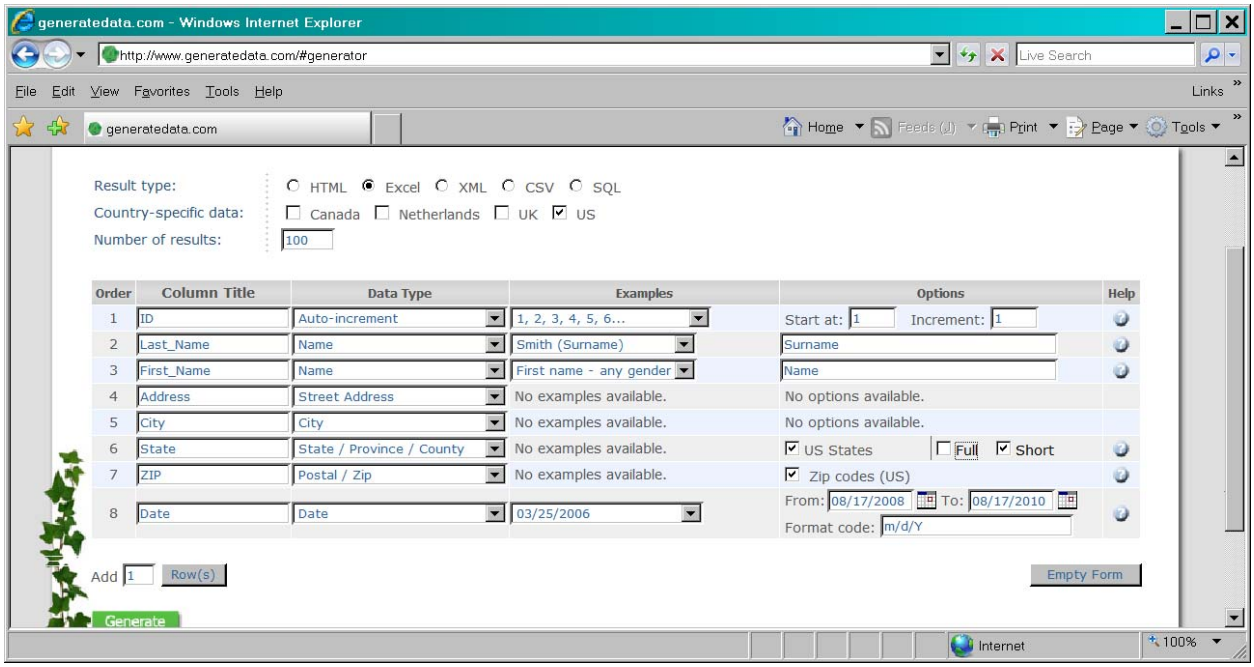
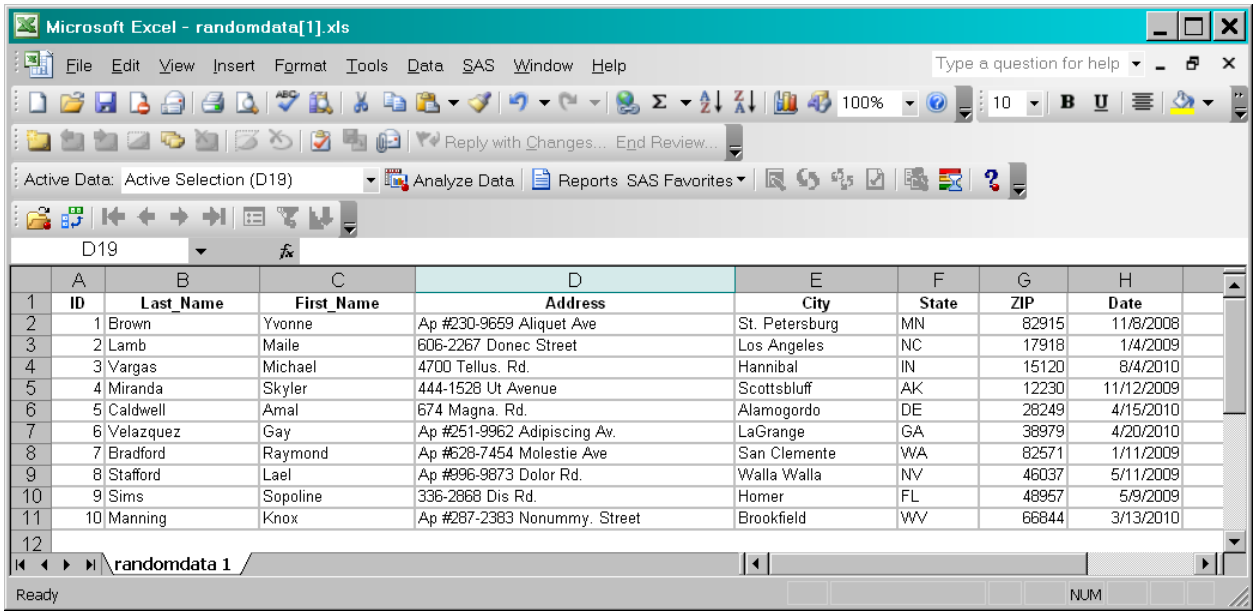



**Figure 1. http://generatedata.com**




| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | ID | Last_Name | First_Name | Address | City | State | ZIP | Date |
| 2 | 1 | Brown | Yvonne | Ap #230-9659 Aliquet Ave | St. Petersburg | MN | 82915 | 11/8/2008 |
| 3 | 2 | Lamb | Maile | 606-2267 Donec Street | Los Angeles | NC | 17918 | 1/4/2009 |
| 4 | 3 | Vargas | Michael | 4700 Tellus. Rd. | Hannibal | IN | 15120 | 8/4/2010 |
| 5 | 4 | Miranda | Skyler | 444-1528 Ut Avenue | Scottsbluff | AK | 12230 | 11/12/2009 |
| 6 | 5 | Caldwell | Amal | 674 Magna. Rd. | Alamogordo | DE | 28249 | 4/15/2010 |
| 7 | 6 | Velazquez | Gay | Ap #251-9962 Adipiscing Av. | LaGrange | GA | 38979 | 4/20/2010 |
| 8 | 7 | Bradford | Raymond | Ap #628-7454 Molestie Ave | San Clemente | WA | 82571 | 1/11/2009 |
| 9 | 8 | Stafford | Lael | Ap #996-9873 Dolor Rd. | Walla Walla | NV | 46037 | 5/11/2009 |
| 10 | 9 | Sims | Sopoline | 336-2868 Dis Rd. | Homer | FL | 48957 | 5/9/2009 |
| 11 | 10 | Manning | Knox | Ap #287-2383 Nonummy. Street | Brookfield | WV | 66844 | 3/13/2010 |
| 12 | | | | | | | | |

**Figure 2. Results from http://generatedata.com**

## EXCEL ADD-IN

This is an Excel Add-In that I found on the internet which helps generate a single table of dummy data.

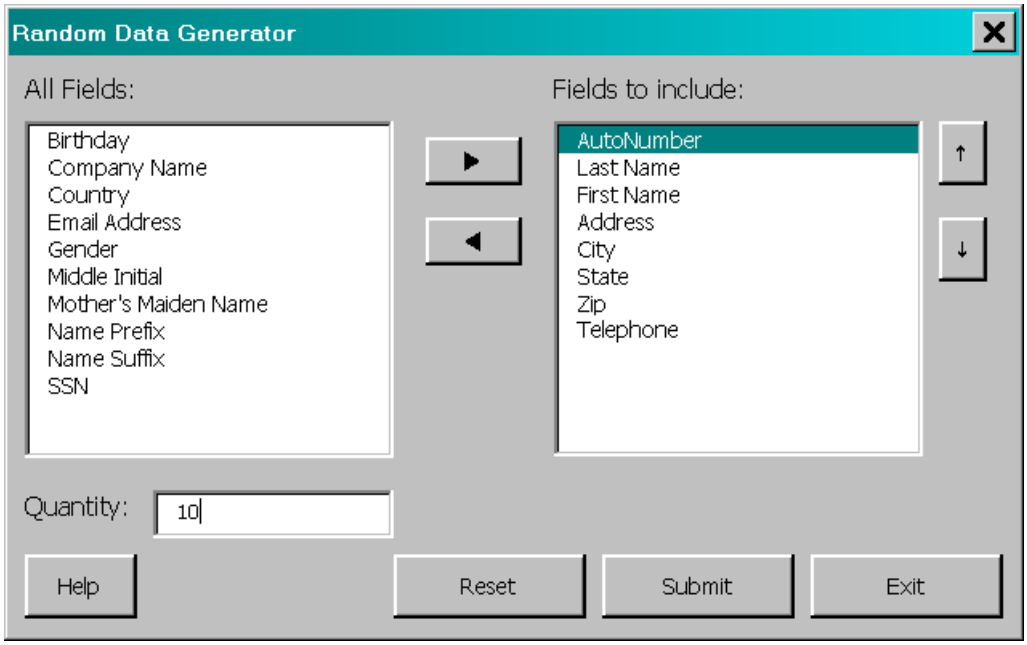
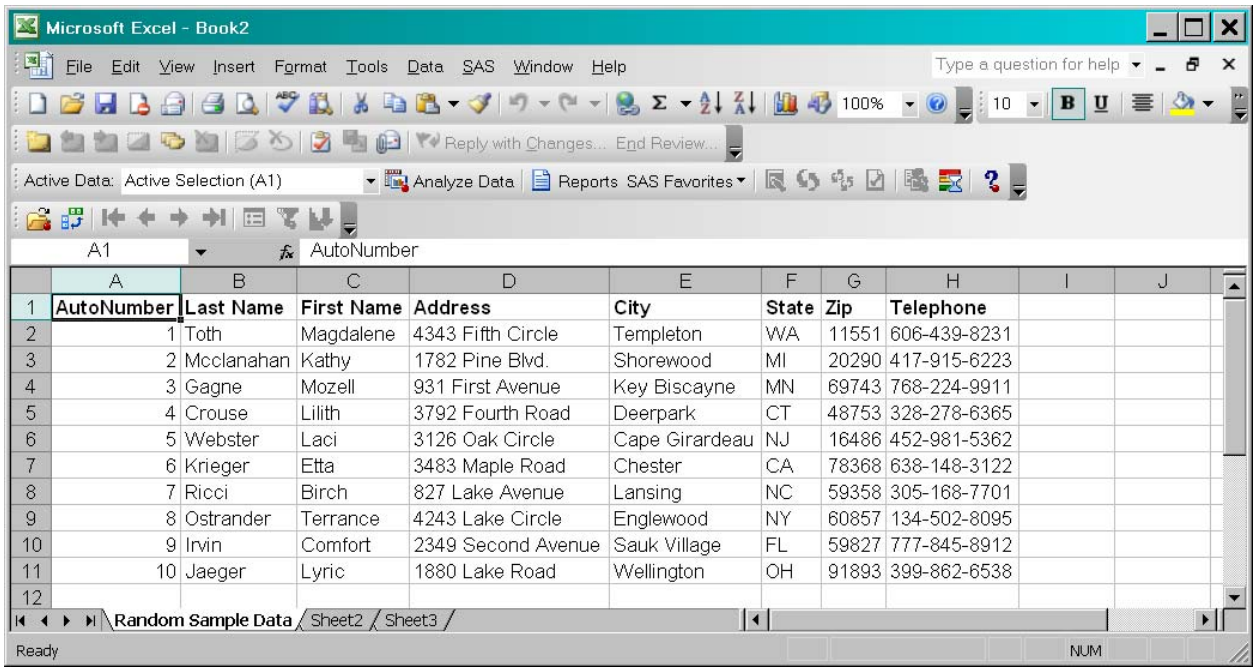



**Figure 3. Excel Add-in from http://datapigtechnologies.com**




| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | AutoNumber | Last Name | First Name | Address | City | State | Zip | Telephone | | |
| 2 | 1 | Toth | Magdalene | 4343 Fifth Circle | Templeton | WA | 11551 | 606-439-8231 | | |
| 3 | 2 | Mcclanahan | Kathy | 1782 Pine Blvd. | Shorewood | MI | 20290 | 417-915-6223 | | |
| 4 | 3 | Gagne | Mozell | 931 First Avenue | Key Biscayne | MN | 69743 | 768-224-9911 | | |
| 5 | 4 | Crouse | Lilith | 3792 Fourth Road | Deerpark | CT | 48753 | 328-278-6365 | | |
| 6 | 5 | Webster | Laci | 3126 Oak Circle | Cape Girardeau | NJ | 16486 | 452-981-5362 | | |
| 7 | 6 | Krieger | Etta | 3483 Maple Road | Chester | CA | 78368 | 638-148-3122 | | |
| 8 | 7 | Ricci | Birch | 827 Lake Avenue | Lansing | NC | 59358 | 305-168-7701 | | |
| 9 | 8 | Ostrander | Terrance | 4243 Lake Circle | Englewood | NY | 60857 | 134-502-8095 | | |
| 10 | 9 | Irvin | Comfort | 2349 Second Avenue | Sauk Village | FL | 59827 | 777-845-8912 | | |
| 11 | 10 | Jaeger | Lyric | 1880 Lake Road | Wellington | OH | 91893 | 399-862-6538 | | |
| 12 | | | | | | | | | | |

**Figure 4. Results from Excel Add-in from http://datapigtechnologies.com**

## USING SAMPLES OF THE ACTUAL DATA

If the data is not sensitive and there is data available, the SURVEYSELECT procedure can be used to generate samples from the actual data. This is typically appropriate if the actual data sources are too large to use for development.

## EXAMPLES

Examples of the different approaches will be provided during the presentation.

## CONCLUSION

Planning for the existence and/or creation of data for development and programming is an important issue to avoid the risk of delays on your project. If actual data is not available at the necessary stage of a project, the creation of dummy data to use for development is essential to build in as a task. The generated data can also be part of the deliverable to be used for future testing and benchmarking.

## REFERENCES

SUGI 22 Paper “Generating Data With the SAS Data Step” by Andrew J. L.

http://www2.sas.com/proceedings/sugi22/CODERS/PAPER74.PDF

Web Based Tool for Generating Sample Data

http://generatedata.com

Excel Add-in for Generating Sample Data

http://datapigtechnologies.com/blog/index.php/creating-sample-data-sucks/

## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Brian Varney, Senior Technical Manager
COMSYS
5220 Lovers Lane
Portage, MI 49002
Phone: 269-553-5185
Fax: 269-553-5101
E-mail: bvarney@comsys.com
Web: www.comsys.com/analytics