Paper 308-2011

Making SAS® Enterprise Guide® Easier to Use

Tom Kari, Statistics Canada, Ottawa, ON

ABSTRACT

Although SAS® Enterprise Guide® is a wonderful tool for data analysis, the design and format of the input data can have a major influence on the efficiency and accuracy of the analytical results. Fortunately, a number of techniques exist both within the SAS® environment and in external data management products to improve the usability of these data. This paper explores the use of views, naming standards, and application of SAS labels and formats to external database tables to demonstrate how typical survey data can be made easy to understand for any level of user.

INTRODUCTION

SAS Enterprise Guide (EG) is a wonderful tool. My friend and colleague Yves Deguire calls it the "Swiss Army Knife of SAS", because it does so many different things. Yes, there might be more focused tools in the SAS toolbox, but I defy anyone to find a better all-purpose tool in the SAS arsenal.[1]

However, the old adage "Garbage In Garbage Out" still applies, and Enterprise Guide is only as good as the data that it can access. This paper will present a number of techniques to make data easier to use for an EG analyst.

The Canadian Census of Population started using Enterprise Guide as a general-purpose analytical tool in 2004. Because of this, the examples in this paper will use our primary Census entities and relationships: persons, who live in dwellings, which are located in municipalities, which aggregate to counties and then to Provinces.[2]

Our data are currently stored in a SQL database environment. While the Canadian Census experience is completely relevant for other database users, most of these suggestions will apply to SAS dataset users as well. If you're using EG to analyze data in Excel spreadsheets, I'm afraid you're on your own!

Broadly, my suggestions for making an EG analyst's life easier falls into these categories:

- Optimize your input data for analysis;
- Use Metadata Server to improve the usability of EG.

OPTIMIZE YOUR INPUT DATA FOR ANALYSIS

The essential goal of these data modification exercises is to present the data to the EG user in a form where it is understandable, unambiguous, and concise.

GET RID OF DATA YOU DON'T NEED

It's not uncommon for a database table to contain fields which will be of no interest to analytical users. In our case, we maintain a number of operational flags and partially-processed variables in the tables, which are essential for managing our processes but are not required for analysis. If these variables are presented to the analytical user, who probably has little or no interest in previous operations, the data will almost certainly be used where they shouldn't be (Murphy's Law, anyone?) or at least generate calls to the support team for clarification. The easiest way to "vanish" these fields is by defining new views on the base tables, and leaving out the unnecessary fields.

PRETTY UP THE DATA YOU DO NEED

Now that you have only the fields you want, tidy them up! Another advantage to using a view to eliminate fields is that those that remain can be i) renamed, and ii) reordered.

Database tables frequently contain field names that are fairly obscure; in our case, a historical data management tool restricted us to four letters for table names and eight for fields, so RO02.R2P1H is perfectly meaningful to an experienced Census analyst. Not so much for a new employee.

[1] Or from any other vendor either, but I may be biased.

[2] As you can imagine, our fully implemented data environment is much more complex, but all of our examples can be illustrated with these easily-understood constructs.

Ideally, the new views should order the fields in a sequence that's meaningful, without worrying about name sort order. That's because EG has options to sort the fields in an input table by physical field sequence or alphabetically by name.

Applying these first two recommendations, we can turn:

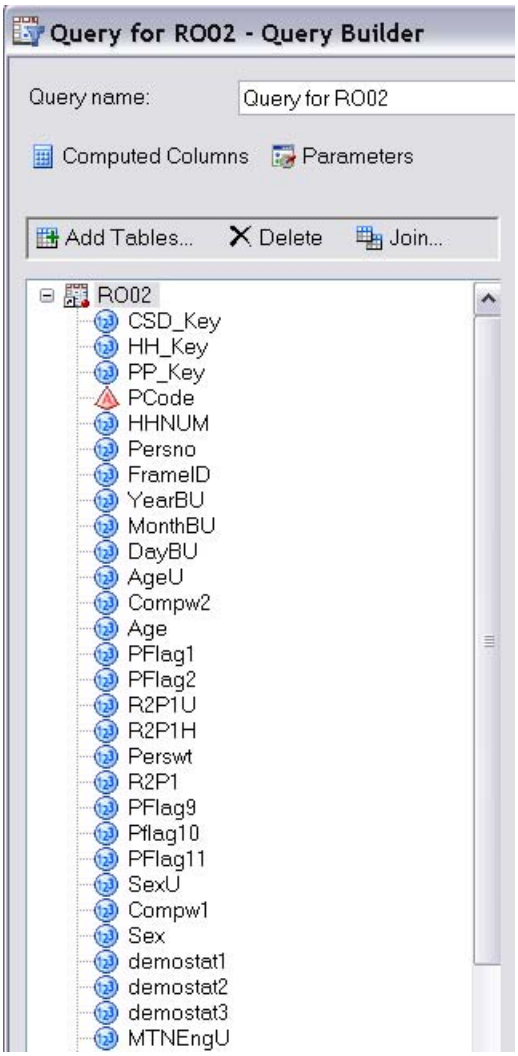

Figure 1. Raw data in EG

into:

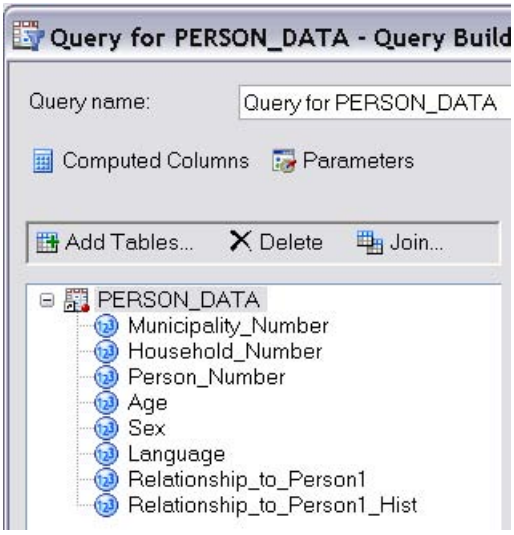

Figure 2. Data reordered and renamed

ADDITIONAL CONSIDERATIONS

Naming standards: Ensure that a consistent naming standard is used. It is particularly important to ensure consistency for weight fields and join keys.

Weights: Input tables may carry a weight field, which specifies how many observations this record represents, and is used in statistical tabulation procedures. It's possible to have multiple weight fields in a base table record. Ideally, trim down your views to only contain one weight field. If this isn't possible, it is essential to have a clear and consistent naming standard for the weight fields, but even then some confusion may result.

Joined tables: It is very common for the statistical variables required for a series of analyses to be contained in multiple database tables, logically related by join keys. If this isn't simplified, analysts will find it difficult to deal with these tables.

1. If possible, design views that bury the details of the joins in database code which won't be visible to the analyst. Expanding on the previous example, we could have:

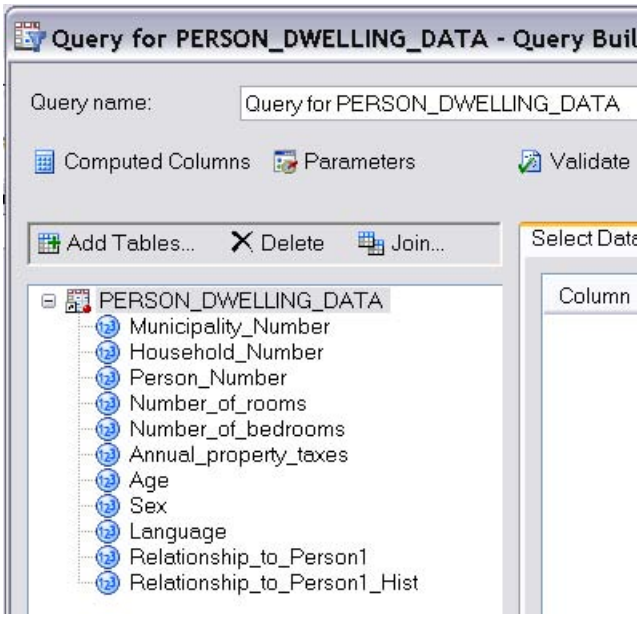

Figure 3. Simplified joined tables

where person data (from the previous example) and dwelling data (number of rooms, bedrooms, property tax) are combined, at a level of individual records for persons (dwelling fields are repeated).

2. Given the fact that views consume no computer resources until they are used, it might be useful to the analyst community to create multiple views, which are really only subsets of one another. In the previous two examples, all of the fields in PERSON_DATA are also in PERSON_DWELLING_DATA, so the latter is really the only view that is required. But if only person level fields are required for an analysis, the analyst might find that using the former is more logical.

3. Canadian Census geographical attributes are hierarchical, so instead of repeating them on dwelling records, they are organized into normalized tables:

```
Province <-->> County <-->> Municipality
<-->> Dwelling
```

with each table containing the relevant fields for that level of geography. By selecting a suitable prefix, again these nested tables can be "flattened" for the analyst, and presented in one view structure. Expanding on the previous example, and adding a weight field, results in:

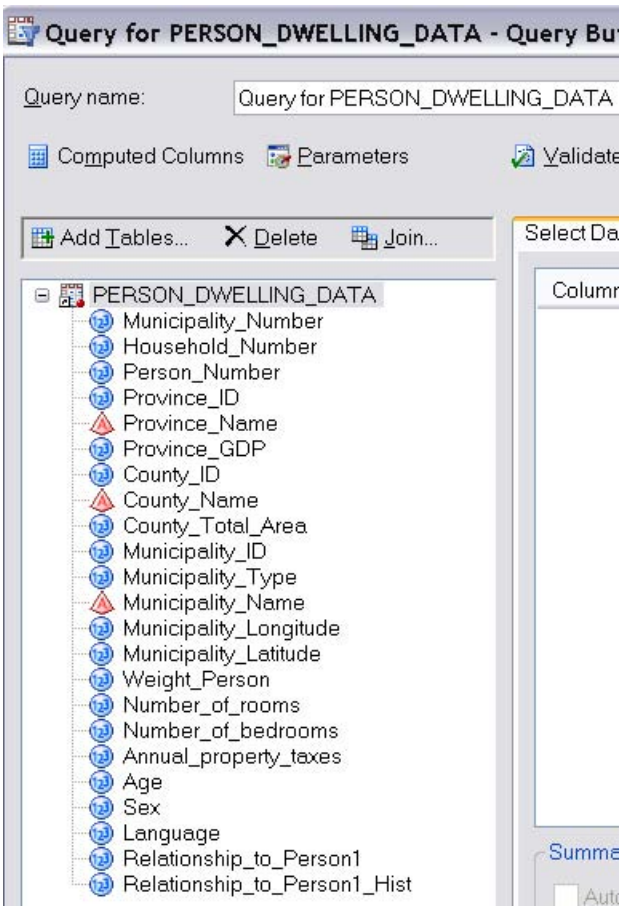

Figure 4. Complex multi table join

which is a very suitable input file for statistical analyses.

4. Finally, if analysts must implement their own table joins, expect problems. These can be reduced by i) ensuring that all of the primary and foreign keys that analysts see are single variables (eliminate composite keys), and ii) never requiring many-to-many joins. Asking analysts to implement either of these database constructs correctly is definitely a bridge too far.

Derived Variables: Enterprise Guide has a very powerful tool (the Advanced Expression Editor) for creating new variables in a query from existing data. However, if data concepts exist that are going to be required by many

analysts, instead of forcing them all to create expressions for the new variables, put them into the view from the beginning. This can be done either by creating new fields in the base table, or by defining new variables in the view, and it will eliminate redundant effort with the attendant risk of error.

USE METADATA SERVER TO IMPROVE THE USABILITY OF ENTERPRISE GUIDE;

For any of you who haven't heard, the 9.2 / 4.2 release of SAS and SAS EG has done away with good old, easy to use EG repository, and replaced it with SAS Metadata Server (MDS). The bad news is that MDS is significantly more complex; the great news is that it contains a TON of features to make the end-user experience of any SAS tool, not just EG, much better.

One great feature of Metadata Server is that it can be used to associate many of the SAS descriptive attributes with both SAS and non-SAS datasets and variables. Two of these that are particularly useful for EG users are labels and formats.

LABELS

Assigning meaningful labels to variables in datasets helps EG analysts in the following places:

- Labels can be selected as the column headers when datasets are displayed in EG;
- The Properties sheet for an input dataset will present the labels;
- In the Query Builder, labels are included in the list of input table variables;
- Labels will be used in many of the statistical analysis and graphing tasks.

Most SAS users are familiar with labels on SAS datasets, but the Metadata Server environment now makes it possible to apply labels to input database files, so that EG analysts see them with no additional effort.

Building on our previous example, the analyst would see the following variable list when designing a query in EG:

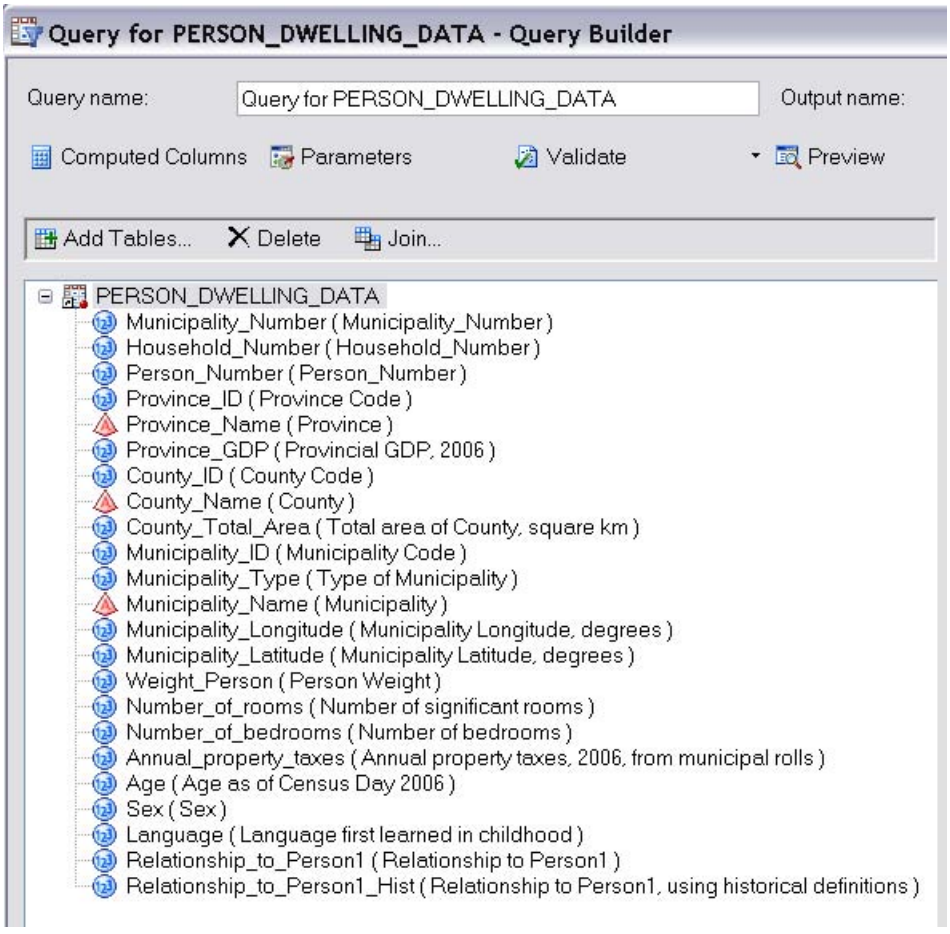

Figure 5. Data in EG with meaningful descriptive labels

For large datasets, applying labels individually using Management Console could get real old, real fast. Fortunately, it is possible to automate the process of applying both labels and formats to the SAS metadata.

FORMATS

Again, the use of Metadata Server allows formats to be associated with non-SAS datasets, in a manner that is transparent to the analyst. As SAS users know, these formats will then be used practically anywhere within the EG environment. This is of great utility to the Census, where most of our characteristic data is coded using numeric variables. Without associating labels and formats with data, dataset display and results of statistical procedures look like this:

PERSON_DWELLING_DATA_SUBSET (read-only)

	Weight_Person	Annual_property_taxes	Sex	Language
1	3.906554306	1568	1	2
2	3.696463918	1420	1	1
3	5.987053371	4402	2	2
4	5.669192724	1242	1	2
5	4.078283666	1086	1	1
6	4.73345209	3442	1	2
7	4.617441008	1141	1	2
8	3.198021804	2682	2	1
9	5.007468966	2402	1	1
10	5.723262617	3744	2	1
11	3.84083881	4400	1	2
12	3.371855771	2279	1	2
13	4.281287705	4245	1	1
14	5.101663372	4496	2	2
15	5.658920309	4866	1	1

	Sex			
	1		2	
	Annual_property_taxes		Annual_property_taxes	
	SumWgt	Mean	SumWgt	Mean
Language				
1	685.12	2994.09	677.14	3036.47
2	390.77	3025.40	319.30	2889.62
4	3.13	3632.00	12.64	3571.87
5	.	.	6.25	4100.00
6	9.10	3674.67	19.87	2737.02
7	.	.	8.97	2028.90
8	.	.	24.33	4281.10
All	1088.12	3012.87	1068.50	3019.46

Figure 6. Data can be hard to understand without formats

Once formats have been associated with the input variables, the same results can be seen to be much more usable:

QUERY_FOR_PERSON_DWELLING_D_0000 (read-only)

	Person Weight	Annual property taxes, 2006, from municipal rolls	Sex	Language first learned in childhood
1	3.906554306	$1,568	Female	French
2	3.696463918	$1,420	Female	English
3	5.987053371	$4,402	Male	French
4	5.669192724	$1,242	Female	French
5	4.078283666	$1,086	Female	English
6	4.73345209	$3,442	Female	French
7	4.617441008	$1,141	Female	French
8	3.198021804	$2,682	Male	English
9	5.007468966	$2,402	Female	English
10	5.723262617	$3,744	Male	English
11	3.84083881	$4,400	Female	French
12	3.371855771	$2,279	Female	French
13	4.281287705	$4,245	Female	English
14	5.101663372	$4,496	Male	French
15	5.658920309	$4,866	Female	English

	Sex			
	Female		Male	
	Annual property taxes, 2006, from municipal rolls		Annual property taxes, 2006, from municipal rolls	
	SumWgt	Mean	SumWgt	Mean
Language first learned in childhood				
English	685.12	2994.09	677.14	3036.47
French	390.77	3025.40	319.30	2889.62
Chinese	3.13	3632.00	12.64	3571.87
Russian	.	.	6.25	4100.00
Spanish	9.10	3674.67	19.87	2737.02
German	.	.	8.97	2028.90
Ukrainian	.	.	24.33	4281.10
All	1088.12	3012.87	1068.50	3019.46

Figure 7. Formats make data and statistical results much clearer

OTHER SUGGESTIONS

MAKE IT FASTER TO TEST QUERIES

Setting up a method to select only a well-designed subset of the data can speed up the process of developing an analysis. If an analyst can develop and test a project using a quantity of data than can be processed in real time, many cycles of trial and error can be run in a short time, before submitting a lengthy production run.

CONCLUSIONS

I hope that the suggestions in this paper will make the life of your analysts a little easier. As the SAS world moves to Enterprise Guide as the default environment for ad-hoc analysis and data exploration, it will be important to wring as many productivity improvements as possible out of the input datasets. Enhancing source data using these suggestions will provide the following benefits:

- Time savings from reducing the number of data transformations required by the analyst community;
- Terminology and codes will be non-technical and appropriate to the analyst;
- The removal of elements that could cause confusion will improve data quality.

REFERENCES

SAS® Institute Inc. 2011, SAS Online Help and Documentation

ACKNOWLEDGMENTS

The SAS Technology Centre at Statistics Canada has proven to be a stimulating, nourishing, and challenging environment. My thanks to my coworkers, particularly Yves DeGuire, and to my clients in the Canadian Census of Population who never quit pushing for more.

CONTACT INFORMATION

I would be delighted to receive comments and respond to questions. Contact the author at:

Tom Kari
Statistics Canada
14th Floor, 100 Tunney's Pasture Driveway
Ottawa Ontario K1A 0T6 Canada
Phone: (613) 951-1352
Fax: (613) 951-0607
E-mail: tom.kari@statcan.gc.ca
Web: www.statcan.gc.ca