Paper 230-28

“I’ll Have What She’s Having” – Serving Up Meta-data to Academic Research Teams

Jeanne Spicer, The Pennsylvania State University, University Park, PA

ABSTRACT

All academics demand reproducibility of results, but most admit that they have trouble reproducing their *own* results, let alone those of another researcher. Continuous experimentation with data defies documentation efforts.

In a large project on the mechanisms of aging, the researcher collecting data on bone density may observe an interesting phenomenon and want to access the data another researcher is collecting on the subjects' blood chemistry to try to explain it. A researcher looking at the subjects' physical activity may want to use another team member's values for body mass index calculated from height & weight and corrected for underlying gender differences. Since there's always a rush to get the data loaded into the database, data entry or calculation errors may be found after analyses have been run. How can members of the team know what data is available, what those data elements represent and be assured that they are all working with the same values in order to produce the same results?

This SAS/Intrnet application attempts to define what is essential metadata for research projects, provides a painless method for researchers to contribute the appropriate documentation and makes it easy for team members to request the data required to reproduce another's results.

INTRODUCTION

This paper looks at problems encountered with managing data for large interdisciplinary research projects with data being collected at multiple universities. Data on animal husbandry may be collected on paper at one university and entered into a SAS file using a SAS/Intrnet application, DNA analysis carried out at a collaborating university may be spit out of laboratory equipment as a spreadsheet, behavioral measures may be entered into an SPSS system file in another building on campus. In order for all of this data to be analyzed together, the files are converted into SAS data tables and managed as one database from the project office. Data, results of analyses, abstracts of journal article submissions and other information are disseminated to all team members from the project website. Much of the website is driven by dynamic SAS/Intrnet applications.

As the research progresses collaborators note that others are getting good results with newly constructed scale scores or other derived measures -- and they want to use those measures in their own analyses. The results published in one paper may warrant further investigation building on those results or trigger inquiry into other related areas. Original data may turn out to have errors or subjects may be eliminated from the study.

Each research team generates notes and documents that serve as metadata: describing the contents of the database, relating data elements and data tables to each other and tracking modifications to the data over time. In order to sift through all this metadata, the project Data Manager needs to collect the metadata with similar care as is devoted to the collection of the substantive data. In order to make sense of all this, new data tables and SAS/Intrnet applications to deliver the metadata to the research team were built into the database. Considerations used in designing this system makeup the subject of this paper. The poster will provide a peek at how the system was actually implemented.

ESSENTIAL METADATA FOR RESEARCH

DOES ALL DATA REQUIRE METADATA?

In a data warehouse for corporate data, the columns and tables structures, while not static, are relatively stable. In research, change is the name of the game. A persnickety data manager can quickly become overwhelmed by the changes and additions to the database researchers require. Attempting to fully document every newly constructed measure is futile. It is better to focus on what is *essential* metadata for the project team members. We concentrate on 3 areas:

1) Edits/corrections to raw data

As data is cleaned and values are scrutinized for validity, there will be edits. Although some cleaning may have already taken place, the data collection sites may not have the resources to look for outliers and crosscheck measures for internal consistency across domains. For example, a laboratory may simply receive a specimen labeled with an ID number and not know the gender of the animal, making it difficult to flag possible out-of-range values. Changes to the data *will* be necessary. Any edit to raw data values must be noted with the change date so that results from preliminary analyses can be corrected.

2) Measures constructed by subject specific experts

In some settings, experts will be called in to construct an appropriate measure. For example, using body weight as a predictor for various motor behavior does not take into account gender differences, bone structure differences, amount of fat in various regions of the body and other factors. An expert on physiology may construct a measure to appropriately compare the physiology of the animals being studied. This would be the recommended measure to use in place of weight or length in other analyses of mobility or longevity. These measures must be documented so that the researchers in behavior can support their findings on activity levels.

3) Files used for research publication

Ultimately, any publication must contain adequate documentation to enable other members of the academic community to reproduce the results reported. Often, 10 o'clock scholars will come up with new measures in time for the publication deadline, but before notifying other members of the team. It is important for the project office to obtain a copy of the file used to produce any runs having results reported and keep it as is (even if an error in the data is uncovered later on.)

MAKE METADATA AN INTEGRAL PART OF YOUR DATABASE

TABLES TO HOLD METADATA

SAS provides metadata[1] at the table and column level in two dictionary tables SASHELP.VTABLE and SASHELP.VCOLUMN. Any SAS application including SAS/Intrnet programs can access the metadata in these tables. If you aren't using them, start now. Metadata in VTABLE includes: Library Name, Member Name, Member Type, Dataset Label, Dataset Type, Date Created, Date

[1] SAS/Warehouse Administrator has other metadata tables that store this information and more. However, due to software and administration costs it is not often possible for an academic research project to take advantage of this product.

Modified, Number of Observations, Observation Length, Number of Variables, Type of Password Protection, Compression Routine, Encryption, Number of Pages, Percent Compression, Reuse Space, Bufsize, Number of Deleted Observations, Type of Indexes, Data Representation, Requirements Vector. Metadata in the VCOLUMN table includes: Library Name, Member Name, Member Type, Column Name, Column Type, Column Length, Column Position, Column Number in Table, Column Label, Column Format, Column Informat, Column Index Type.

Because these table views are created dynamically by SAS and do not store historical information, the project data manager should not be afraid to design additional metadata tables to supplement the information contained in the VTABLE and VCOLUMN tables.

You will want a table to store **table-level** metadata linked to the VTABLE by Library Name and Member Name fields. This table includes a column to hold text data similar to a ‘memo’ field and a column to identify a point of contact for questions on the file. It also holds various flags, for example whether the table is to be included in one of the dynamic summary reports that can be requested from the website. Another useful field is a subject tag used to relate tables dealing with the same subject area, for example several labs may conduct analyses of blood samples, each generating their own data table. All tables can be given the same subject tag “Blood Chemistry” making it easy for the researchers to find all the appropriate tables to that subject area.

The table level metadata is also used to construct the text frame for the website’s data download and that uses the SAS/Intrnet Xplore application. The text frame provides table descriptions for all tables that appear in Xplore’s data tree.

A **column-level** table with the key fields of Library Name, Member Name and Column Name for linking to the VCOLUMN table is created. This table would also have a ‘memo’ field for brief descriptions of calculations used in the construction of the variable. Edits to correct data can also be noted and dated.

You may require a **subject-ID-level** table to record changes that apply across the database to a particular person such as the withdrawal of a subject from the study.

DEVELOP A PROTOCOL FOR COLLECTING METADATA

Each of the three categories of data requiring metadata needs to have a protocol established for validation of the data values and their documentation.

For edits/changes, the person noting the error must submit notification to the data manager. Typos may require nothing further than correction and notation in the column-level metadata table with the date of correction. Suspicious outliers however may need to undergo further review by a member(s) of the team to confirm whether the data should be corrected or simply flagged as being extreme values.

For newly constructed or recoded measures, the data manager and others will want to examine the data for accuracy and review the documentation to make sure that results can be replicated using the data currently in the database.

For files used to generate published results, the Principle Investigator will need to ensure that all data and supporting documents are submitted to the project office at the time of publication. If not, files have a tendency to ‘evolve’ as the researcher experiments with modifications to their original calculations. Eventually the file no longer contains the values used to generate published results.

The Data Manager can facilitate the data and metadata submission process by providing a web-based application enabling a researcher to upload files and documents to the project office. Include a link from the project website to a metadata collection page that is as easy to use as email. The page should provide plenty of room for composing a text message and allow for attachments or access to ftp. Be sure to capture some basic information about the table, variable or subject id and immediately update the metadata tables. For example, responses to these fields allow you to at least flag ID 3546 in the subject-ID-level table:

Metadata report for:
○ Table
○ Column
● Subject ID

Name: 3546

Type:
○ Edit
○ Recode
○ Calculated
● Delete

The application should automatically notify members of the team who serve to review and verify data.

NAME AND LABELS PROVIDE METADATA TOO

From time immemorial, or at least since COBOL was developed, consistent naming conventions for variable names have been touted as the best possible documentation. And it's true, the names I give my own variables are easy to remember and are self-documenting -- to me! Researchers contributing data to the database may be collecting the data with software that limits names to 8 characters, forcing unintelligible abbreviation. Software may have loose requirements for formatting of column headings allowing internal spaces and special characters that get stripped or substituted in conversion to a SAS table. People in various subject specialties have different terminology for the same thing. The Data Manager will be responsible to make sure that all columns submitted to the database have unique and reasonable names and they will be damned for it. I have found that the best thing is to impose naming convention standards to the variable LABEL rather than the variable NAME. When it is necessary to change names given by the researchers to their columns be sure the label allows the researcher to match their column names to the new ones created.

For example, a researcher contributes an Excel file pertaining to weigh-ins during treatment. The spreadsheet has these column headings for each treatment date:

"Date" "weight_oz" “wgt2”

Another researcher sends a file pertaining to the subjects admission into the program and related weights. The variable names include:

“Date” “Date1” “1st weight” “Date 2” “Final wgt”

The data manager can eliminate duplication of names for different dates and provide more information to distinguish between the various weights by the use of creative naming and labeling.

```
name: treatment_date label: "Date of Treatment"
name: wgt_oz _pre  label: "Weight in ounces, pre-treatment"
name: wgt_oz_post label: "Weight in ounces,  post-treatment"

name:  birth_date    label: “Date of Birth”
name:  admission_date label: “Date of Admission”
name:  wgt_initial   label: “Weight on admission”
name:  release_date            label: “Date of Release”
```

name: wgt_final label: “Weight on release”

name: treatment_age label: “Age (days) at Treatment CALC: treatment_date - birth_date”

The new variable names standardize the abbreviation for “Weight” and “ounces”. Using the word “Weight” in the label allows for keyword searching of the label for any variable dealing with the subject’s weight or any variable pertaining to “pre-treatment” measures. A new calculated variable “Treatment_age” is constructed at the project office and included in the data set. “CALC” indicates that these values were calculated, not entered. And the label includes the names of the variables used in the calculation.

DOCUMENTS ACT AS A ‘SAFETY NET’ FOR METADATA,

Don’t let documentation fall through because you are swamped or the documentation provided is obscure or verbose (never happens with academics!) Develop a fail-safe method to flag the files & columns that require further explanation. This can be a simple link in the data table level metadata file to a Word document, web page or email message. If you can’t individually document each column in a file, at least link to table level documentation.

PROVIDE MULTIPLE, INTEGRATED ACCESS METHODS

The project website is navigated via a Table of Contents frame on the homepage. Each subject domain has a link to its own page. One important link on the subject domain page is the ability to list all tables (and their contents) related to the subject area using the subject tag column in the column-level metadata table.

The website’s Table of Contents has a link to a Column Lookup query page. Information on columns can be retrieved by keyword in the name or label including notes in the column level metadata table. The query can be limited to columns in a specific table or spread across all tables in the database to allow the researcher to locate for example any column providing information on the subject’s “weight”.

The Table of Contents also links to a Subject ID Lookup query page. General demographic information on each subject can be displayed along with any metadata in the subject-ID-level metadata table.

Researchers who know that a particular file was created for a publication or experiment can retrieve it by “Investigator/Publication Name” or in a sense be able to ask: “Do you have a copy of that file Sally used to compare different ways to calculate body mass index--you know the stuff she presented at that conference?”

The Table of Contents has a link to a Database Status summary report that dynamically provides lists and counts of tables, variables, modification dates, etc. Users can drill-down to specific table and column metadata from the report itself.

Macros to display table, column and subject level metadata were written and are called from multiple places on the website. This minimizes coding and provides a consistent ‘look’ to the metadata displayed.

MULTIPLE POINTS OF ENTRY

Good metadata tables allow for any number of access strategies to be implemented easily.

Be able to query metadata by:

- keyword
- change status
- publication/study name
- file source

INTEGRATE ENTRY POINTS THROUGHOUT YOUR WEBSITE

Provide ability to query the metadata tables from multiple entry points on your website:

- Domain specific descriptive pages
- Master database status report
- Metadata query page
- Xplore application text frame

REWRITE PROC CONTENTS

Design a good method to report metadata that can be re-used to generate the output report from all search strategies. Rewrite PROC CONTENTS to include:

- Text of table-level documentation
- Text of column level documentation
- Link to table-level documentation as a ‘safety-net‘ anytime metadata for a column is requested, but not provided in the column-level documentation

ACKNOWLEDGMENTS

Much of the work for this paper was in support of the “QTL Analysis of Age-Related Phenotypes” project conducted by the Center for Developmental & Health Genetics, Penn State University and funded by the National Institute on Aging.

CONTACT INFORMATION

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

Jeanne Spicer
Penn State University
813 Oswald Tower
University Park, PA 16802
Work Phone: (814) 863-8321
Email: spicer@pop.psu.edu