How to add metadata to your data so that you and others can make sense of it

Dr Richard R. Plant 20120220

JISC DMSPpsych Project

Introduction

In the course of your working life as a professional Psychologist you will routinely generate data. By now you will have already generated varying types and amounts. Some will be quantitative and some will be qualitative, some might be from traditional experiments, observational studies or clinical work.

The common thread that links the data you generate is that it's often rendered unusable in a frighteningly short amount of time. This is irrespective of the perceived importance at the time you collected it, did an analysis and wrote it up!

Data can become unusable for a variety of reasons, e.g. data losses through hard drive failure, accidental deletion, computer upgrades, software that can read the data becomes obsolete or file formats change. However the most common reason is that you, as the person who generated the data in the first place, forget what it means! Or rather you forget how the data was coded which means you can no longer make sense of it. This is especially true if you are looking at the data in isolation without the written interpretation or publication that went alongside it.

With the big push to reuse and share data more widely this is becoming of paramount importance. If no one can understand your data it will be devalued. Remember it might be you a few years down the line that wants to reuse the data you collect today!

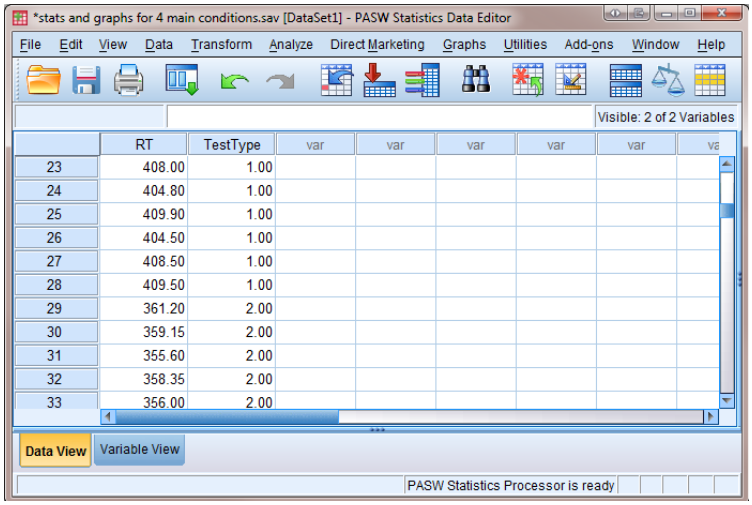
By describing your data with metadata you effectively annotate it so that you give it context and meaning. Metadata to all extents and purposes is data about data. Doing this isn't as complicated as it sounds and it's better to do it as you go along rather than see it as an additional onerous chore that's done at the end of a project.

For DClinPsy trainees it is now a mandatory requirement that they annotate their data with metadata. This is because supervisors will now be checking actual data files and may want to rerun analyses that have been carried out by trainees.

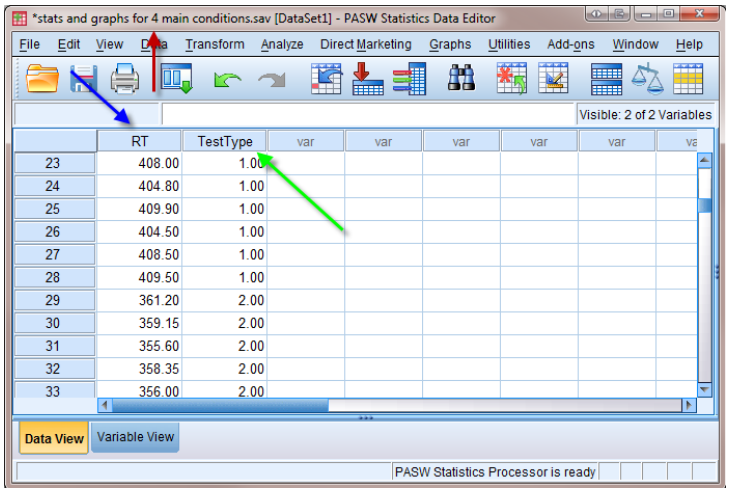
Choosing sensible filenames

On many occasions it's the simple things you can do that make your data much more understandable. For example, you can start by using meaningful file and variable names.

To help illustrate how you can add simple metadata to make your data more understandable I'm going to show you how to make some classical quantitative data more understandable. The central concepts will apply to your data even if there is not an exact match. If we look at the SPSS screen grab below we can see a fairly standard layout. Can you spot any problems that might prevent the data being understandable in 12 months time?



Hopefully you've spotted there are three immediate problems? The first is that the filename shown with the red arrow is meaningless. Second, you could guess that "RT" (blue arrow) means Reaction Time or does it mean Relative Therapy or something else with the acronym RT? Is the measure seconds, milliseconds, a score on a psychometric test? Thirdly, as for "TestType" (green arrow) we can see that there are two or more conditions but what type of test, what are the conditions?

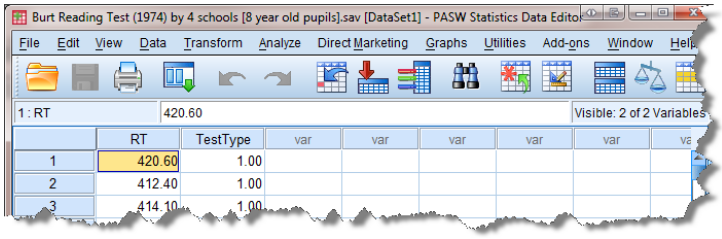


As a starting point a sensible file name can help orientate us as to the contents of any file. This is especially true of a data file. So perhaps we could make it clearer by renaming the file from:

stats and graphs for 4 main conditions.sav

to

Burt Reading Test (1974) by 4 schools [8 year old pupils].sav

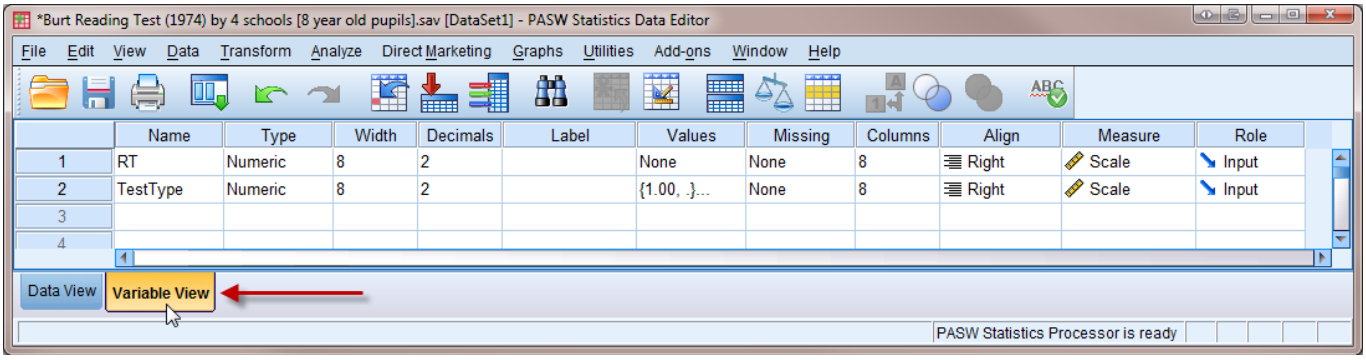


Instantly you can now see that RT actually means the Burt "Reading Test" scores on the 1974 revision of the test and that the four conditions were actually because the test was administered in four different schools with groups of 8 year old children. You could also infer that the point of the study was to look at language as the BRT is concerned with assessing phonetics usage. Hopefully this illustration has given you a eureka moment!

Using the SPSS Variable View to help describe your data in a more meaningful way

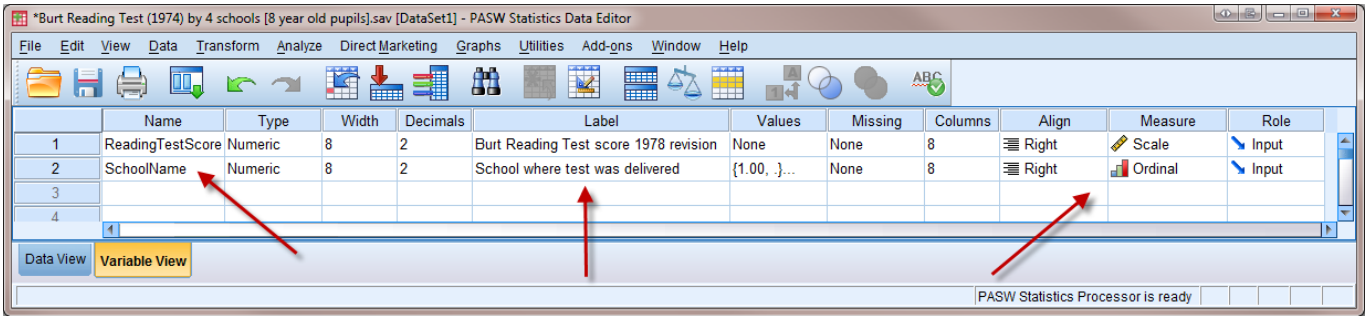
SPSS in common with many other statistical analysis packages has features that let you describe your variables in a way that helps clarify their purpose and how they might be used. In SPSS you do this in the "Variable View" tab.

This process is technically known as adding annotation metadata. Simply adding additional textual data which explains the variables and what they hold.



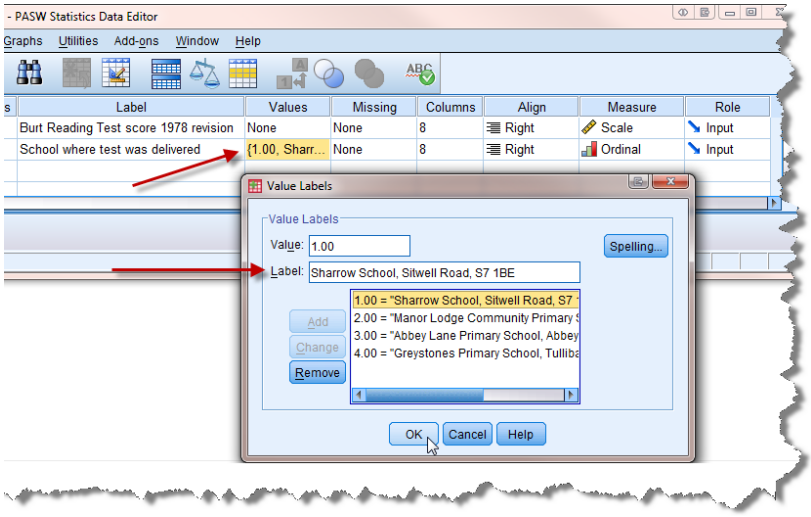
In the example above we can see the two variables discussed earlier. The first the Burt Reading Test score and the second is which school the test was delivered in. To begin with it's helpful to set the type of measure and then actually choose a sensible name and label for each variable.

Now that the variables have been given longer more meaningful names we can see how the context has become clearer. By doing this it would be possible for your supervisor or a third party researcher to reanalyse your data. In short your data makes sense outside of your written thesis. Describing your data sensibly takes very little additional effort. The best way to approach this is to try and think how you would want the data labelled if you were to revisit it in five years time. Could you understand the data if you opened up the SPSS file alone?

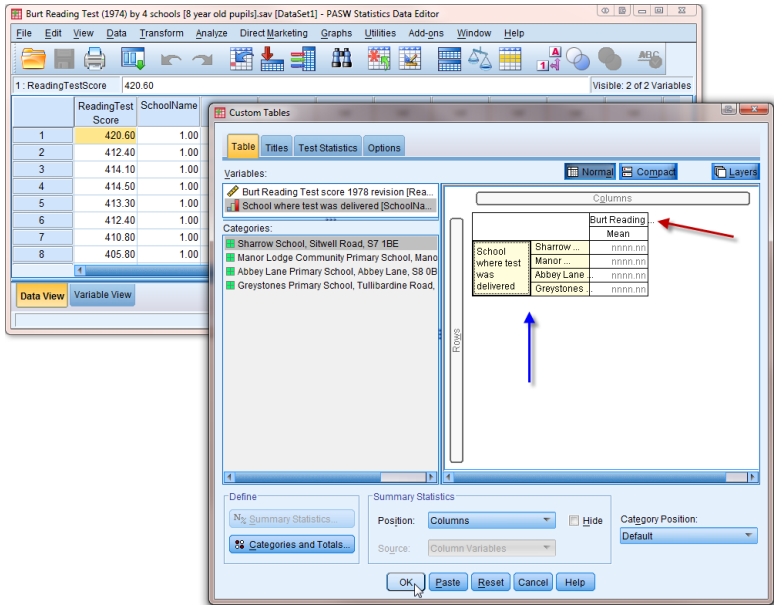


For ordinal variables, in this case where the test was administered, you should also use descriptive labels to help clarify what each category means.

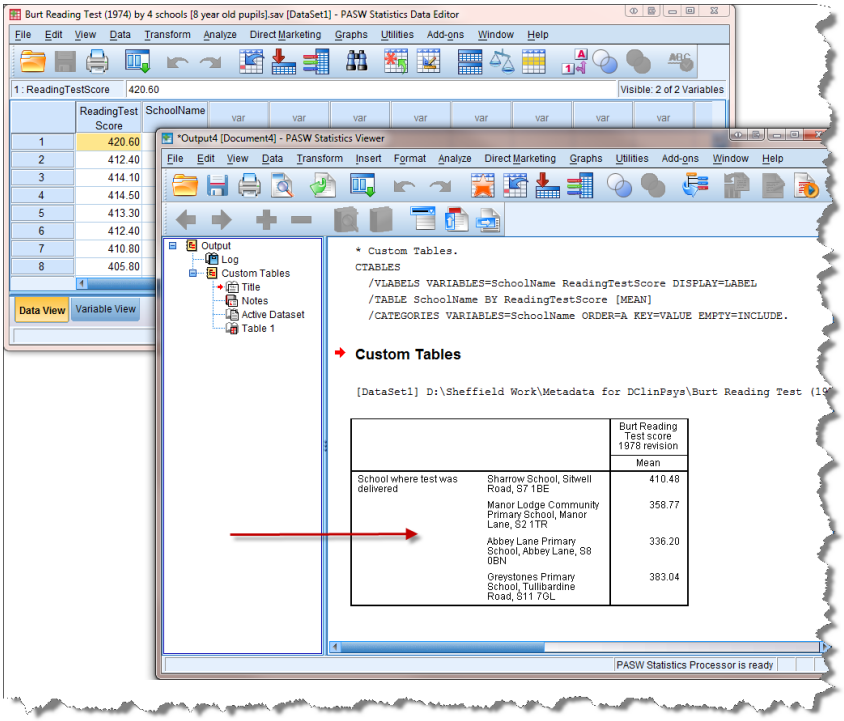
In the example shown below I've added the name of the school, the road it is on and the postcode for each value. If I'd just left the values without sensible labels it would be impossible to know which schools were involved in the study. Even with the accompanying thesis, which would probably describe the schools who took part in some depth, you would still not know which schools data was which. Often at the time you carry out your statistical analysis such variable names may seem meaningful to you as you are intimately involved with the research. Again you need to put yourself five years in the future. Would it be as clear then? If not, you need to think about providing additional descriptions and contextual information.

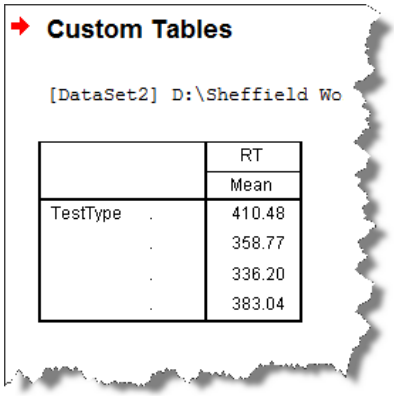


When we come to carry out an analysis the benefits of taking the time to fully describe variables and categories becomes clear as can be seen from the Custom Table analysis shown below.

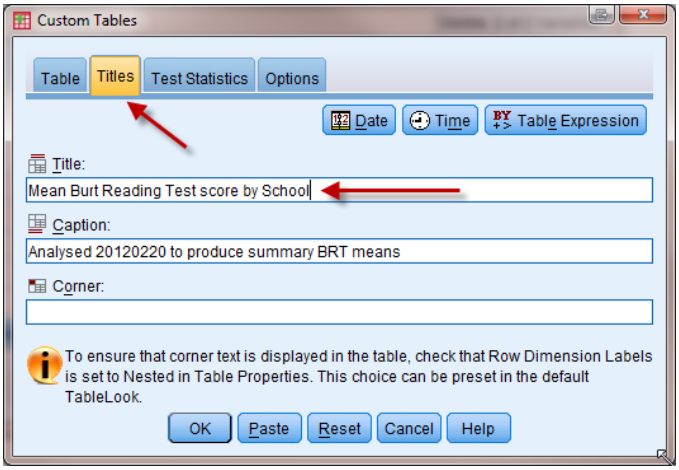


When we actually carry out the analysis the resulting table is extremely clear. Contrast this with the same analysis carried out on the first version of the file with poorly described data. Obviously this is an extreme example but it illustrates how easily meaning and context can be lost.





Even though you may decide not copy and paste the results table directly from SPSS into your thesis it's still sensible to give tables proper tiles and descriptions. In most SPSS dialogs there is a tab where you can choose to enter titles and other helpful contextualising information.

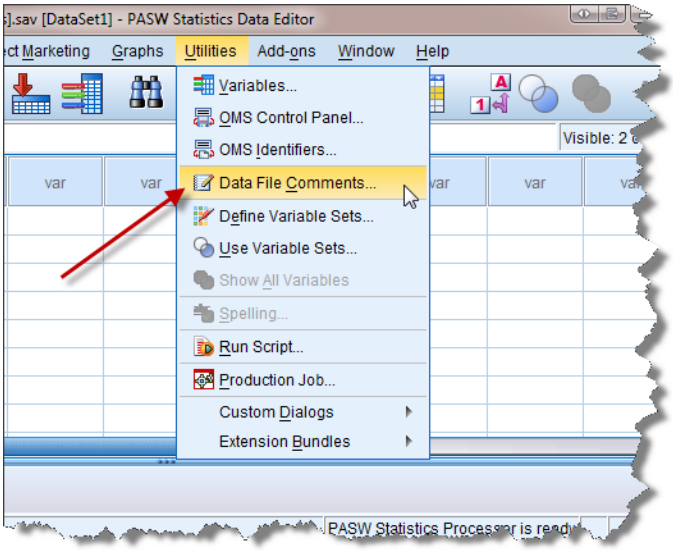


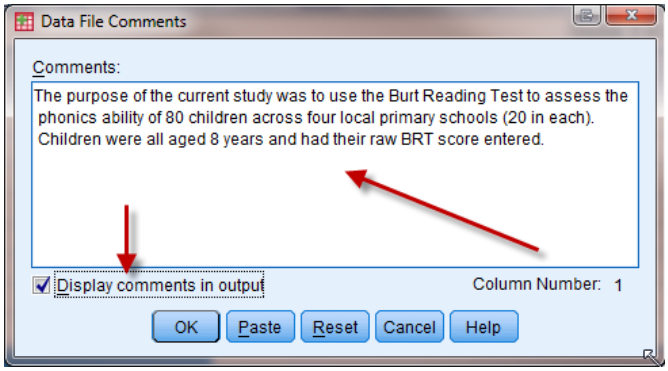
Mean Burt Reading Test score by School

		Burt Reading Test score 1978 revision
		Mean
School where test was delivered	Sharrow School, Sitwell Road, S7 1BE	410.48
	Manor Lodge Community Primary School, Manor Lane, S2 1TR	358.77
	Abbey Lane Primary School, Abbey Lane, S8 0BN	336.20
	Greystones Primary School, Tullibardine Road, S11 7GL	383.04

Analysed 20120220 to produce summary BRT means

Finally you have the option to leave "Data File Comments" in SPSS. These are short text based notes to yourself or others that help orientate them to the context and purpose of the data file. To add comments select "Data File Comments" from the "Utilities" menu.





Ideally you should describe the purpose of the current study data as succinctly as possible with the goal of giving the reader an insight into the reason why it was collected.

Adding administrative metadata

Unlike annotation metadata which explains the data itself administrative metadata describes who created the data, the title of the dataset, the format the data is in, when it was created etc.

Normally you would create this as a separate text file. That is, a plain text file that can be opened in a simple text editor like notepad. This is because the goal for this file is that it should be able to be read as easily as possible even if the software used to create the data is no longer available. For example, SPSS Inc could have gone out of business and the person who might want access to the data needs to make a judgement call as to whether it is worth trying to retrieve the data which is in the proprietary SPSS file format.

The most common way of writing administrative metadata is to use the "Dublin Core DCMI Administrative Metadata" format. This is simply a set of "elements", or sections, that you should aim to complete. A full list of the 15 elements is shown below.

Element Name	Element Description
Title	The title of the dataset
Author/Creator	Who created the dataset
Subject and Keywords	One or more keywords that describe the dataset
Description	A description of the dataset
Publisher	Who published the dataset
Other Contributor	Anyone else or other organisation who helped create the dataset
Date	The date the dataset was created
Resource Type	What type of resource is the dataset
Format	The type of format the dataset is in
Resource Identifier	A unique identifier or URL to the dataset
Source	If from a paper version list the source document
Language	What language is the dataset in
Relation	Is the dataset related to another or another study
Coverage	Is there some geographical coverage or some other limiting coverage factor
Rights Management	Is there a formal rights management policy

Dublin core metadata is a lot simpler to use than it sounds and once you have used it previously you can setup a boilerplate that you can reuse for other datasets. The easiest way to conceptualise its use is to liken it to a catalogue record for a library book.

An example administrative metadata file is shown below for the Burt Reading Test phonetics discussed earlier.

readme.txt

Tile = A study to compare four local schools reading age on the Burt Reading Test
Author/Creator = Dr Richard R. Plant
Subject and Keywords = Psychology, Burt Reading Test, phonics, phonetics, reading age
Description = This study used the Burt Reading Test to evaluate the reading age of children aged who attended four primary schools in Sheffield. A total of 80 children participated (20 per school) and were of roughly equal gender split.
Publisher = The University of Sheffield, Dept of Psychology
Other Contributor = None
Date = 2012-02-20
Resource Type = SPSS Inc, SPSS 18 data file
Format = SAV/data file
Resource Identifier = http://etheses.whiterose.ac.uk/9765341/
Source = Original version
Language = en-GB
Relation = IsRelatedTo http://www.syntheticphonics.com/burtreadingtestpage.htm
Coverage = Sheffield, GB
Rights Management = http://www.shef.ac.uk/library/services/copyetheses

By creating both annotation metadata and administrative metadata you are helping ensure that your data remains as usable as possible and that you are clearly identified as the author and receive appropriate credit for it.