Block 1 - From questionnaire to data file

1.2.4 Second look at data from a major survey [Updated 1 September 2010]

Exemplar: British Social Attitudes (1989)[1]

The raw data for the 1989 British Social Attitudes survey were initially supplied by the UK Data Archive in a ***.dat** file which, on my previous computer, Windows interpreted as a Wordperfect file. Wordperfect displays the data in **Times New Roman proportional** font which is not only horrible to look at, but also, because the columns are not vertically aligned, is impossible to interpret visually.

```
2382901090128291912O3      511201  2  0201150450201        1    61  00.6666
2382902220381511042210 311  1111121221  13113112112121132 13343132  031 3943122
23829034            1006  33312    2    42 090106122 222 2112114410390903
2382904     152311151        3
2382905
2382906
2382907                        0301141133311421 22 113112     232344
2382908010407 11131210       1212112312     231211412111111
2382909
2382910
2382911412 4            2 12 4444  3 3
23829128  8  3831       1           2                       14122
2382913   1330342    1 1 1222  2212211211 226                       2222227
2382914102251 15911                                        1
2382915 22      21  2       12                      069010907329604
2382916    000010341 22222203    071030808327108    00061021111           1
23829172211     4       130611112 22122123045919125O789                  3
2382918
2382919432221423223 1     43122312223   21    3 2122111  2223 1       333422224
2382920223535525211                              32221122221232211511212
2382921134133312233 31                              342232222222222223 33
2382922                          1      335333        2344444441311
238292323221  22322333323    225078921241064115      010265442434100837 3215
```

Wordperfect is not installed on my current machine and Windows thinks the *.dat file is a movie!!

Although SPSS can read the data in this format, provided the file name is enclosed in primes (apostrophes) i.e. **'myclass.dat'**, it is always useful to be able to inspect the original data visually.

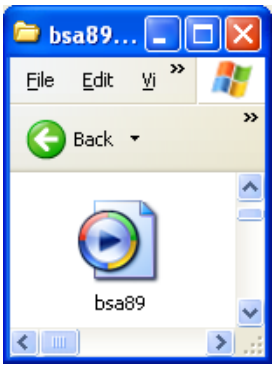
For our purposes the data need to be displayed in a **fixed-width** font to make it easier to find our way around inside the file. On my previous machine I could change the font to **Courier New** inside Wordperfect: on this machine I had to open the file with Word and convert it.

You don't have to do this yourself, but you may find it useful to follow the steps in case you have to do something similar in future. In any case we shall be using these data in later exercises and there's a little test at the end to see if you've understood how the data relate to the questionnaire.

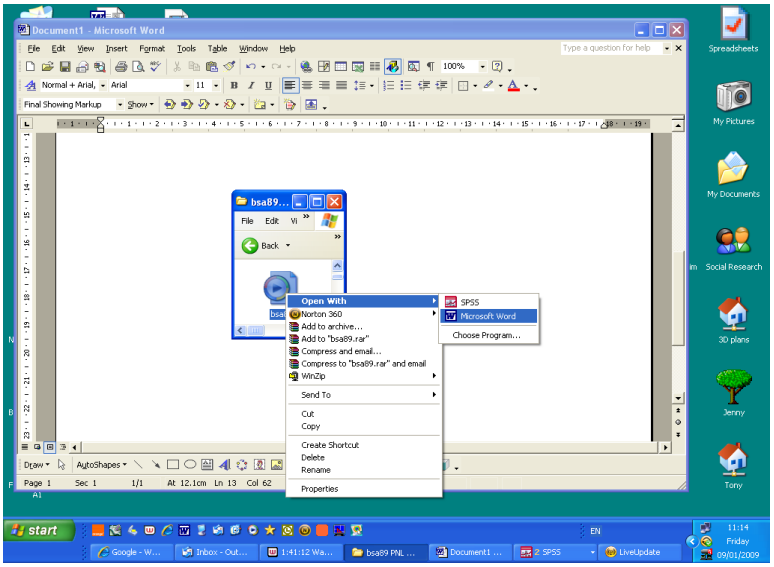
The original data file was supplied as a WordPerfect file **GB8901.DAT**, but I renamed it **bsa89.dat** and then converted it from **Times New Roman proportional** to **Courier New fixed-width** font in ***.txt** format in file bsa89.txt[2]

[1] For data details see: http://www.data-archive.ac.uk/findingData/snDescription.asp?sn=2723

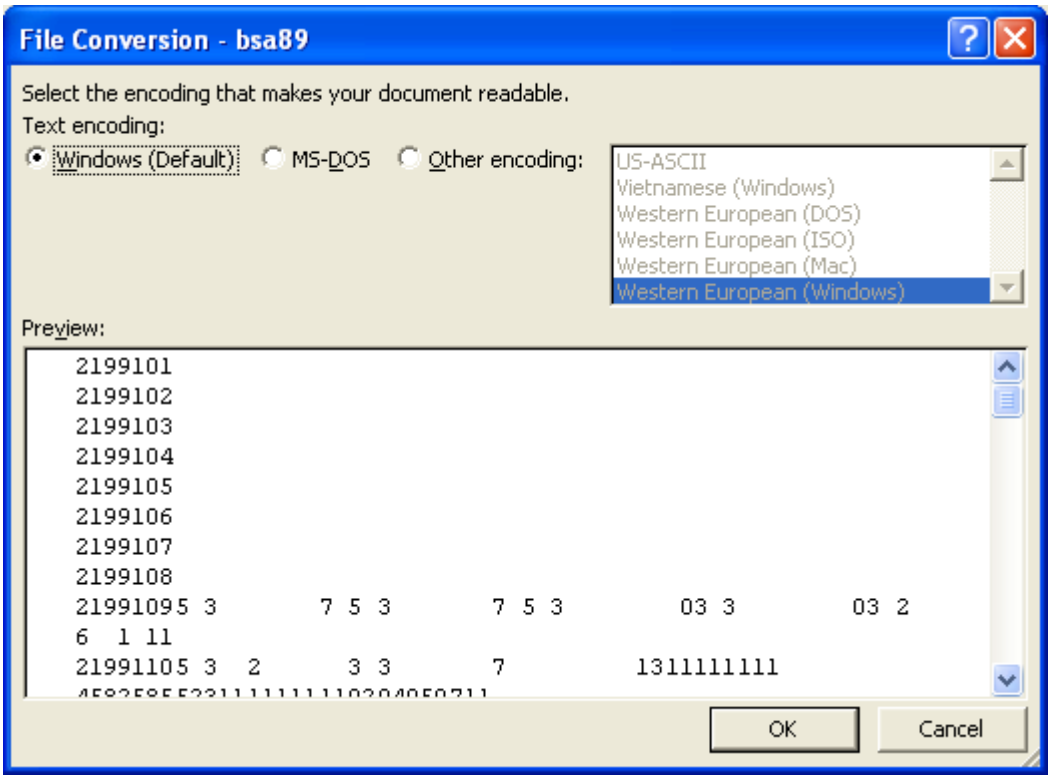
[2] You can also download the original raw data file bsa89.txt (3.9 mb) **from this site**



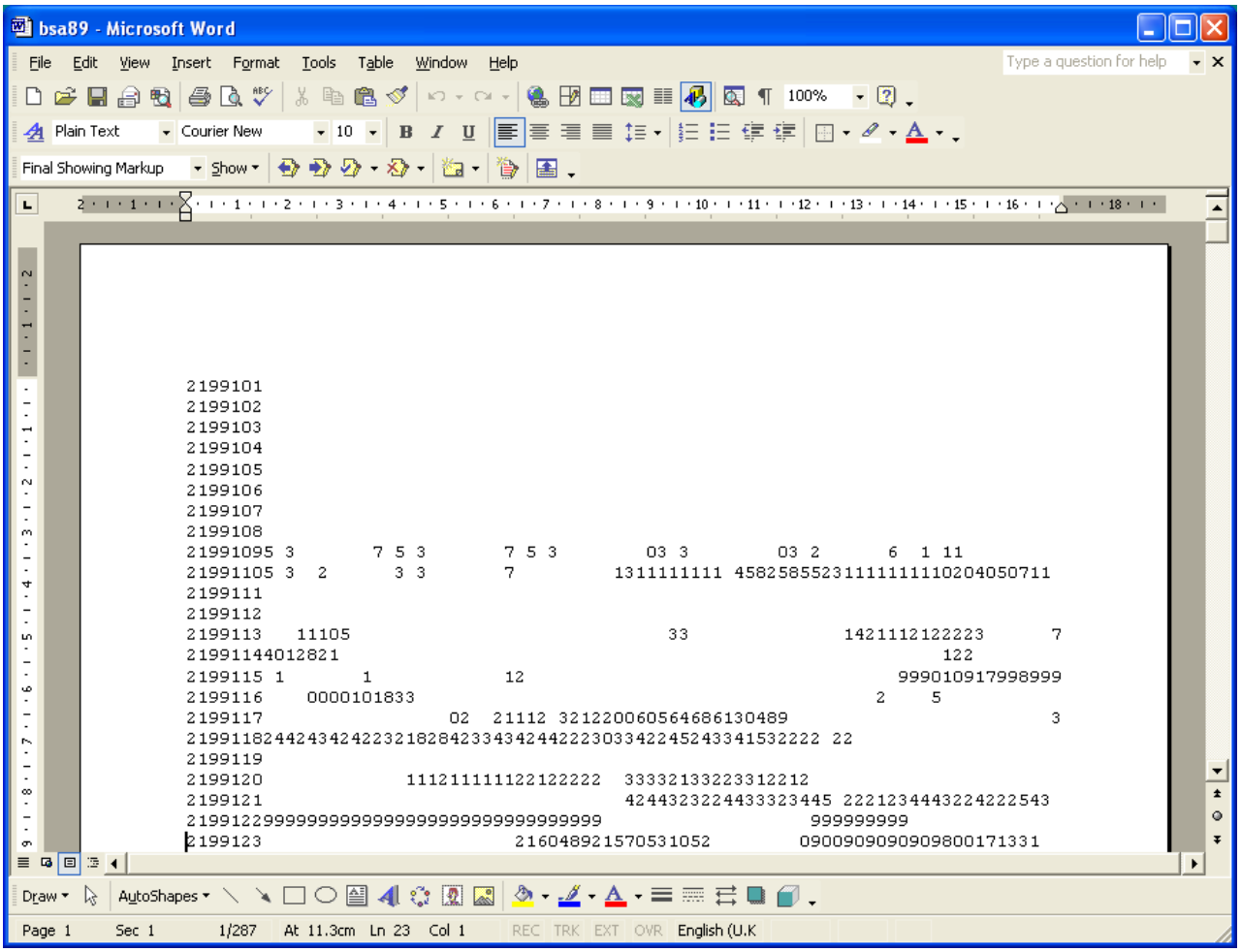
Right click the icon, click on **Open With**:



… then on **Microsoft Word**, which displays:



Press **OK** and following will appear:



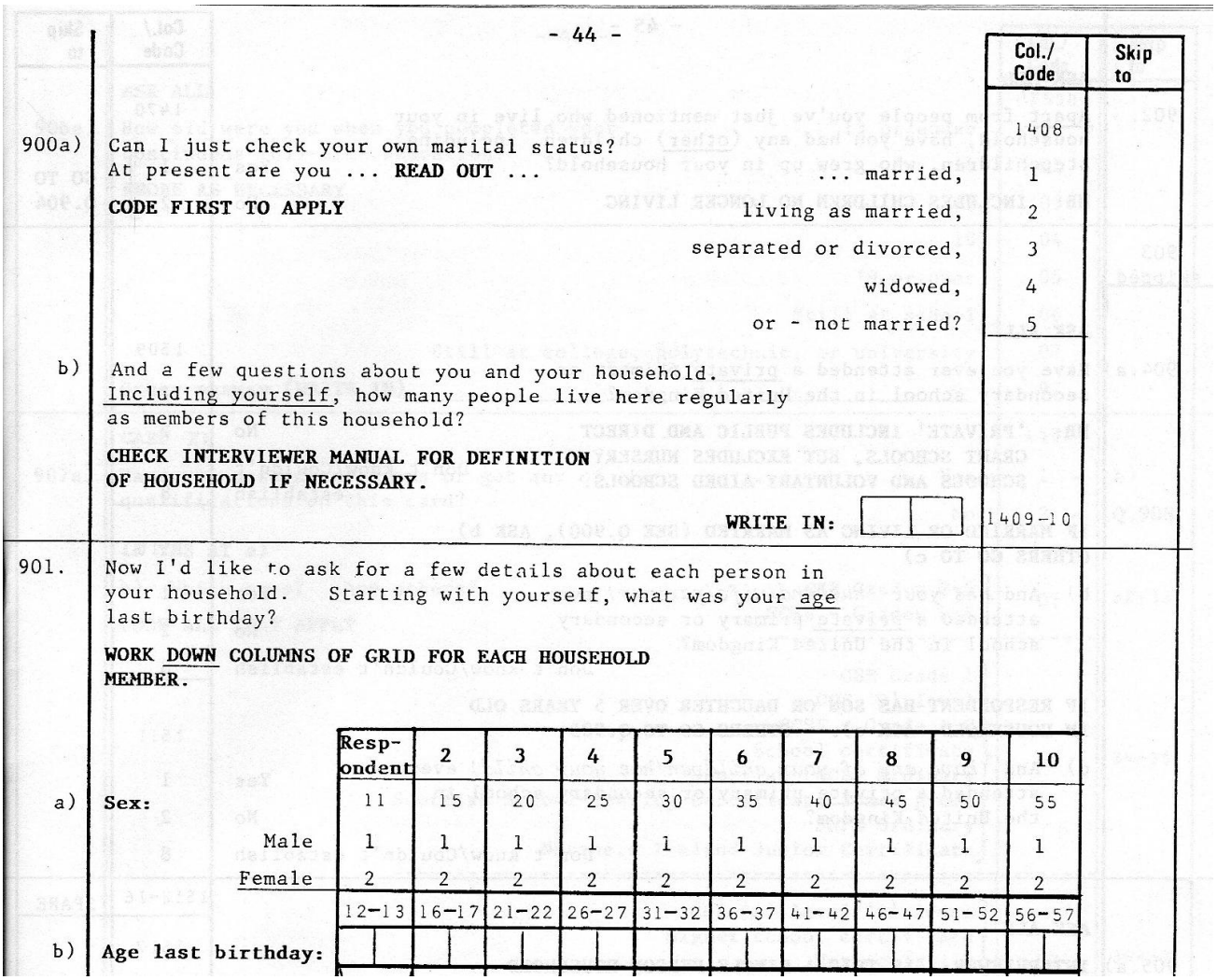
This is a conversion of the original data from **Times New Roman** **variable-width** font to plain text in **`Courier New`** **fixed-width** font. The data are displayed in 80-column lines in which all 80 columns are vertically aligned. The 1989 British Social Attitudes survey has 3024 cases and 23 lines (records) per case (that's a lot of lines!).

The first case is displayed in the screenshot above.

Second exercise on a major survey

British Social Attitudes 1989 – Version B, questionnaire page 44

Below is a facsimile of part of the questionnaire dealing with household information:



- 44 -

		Col./ Code	Skip to
900a)	Can I just check your own marital status? At present are you ... **READ OUT** ... **CODE FIRST TO APPLY**	1408	
	... married,	1	
	living as married,	2	
	separated or divorced,	3	
	widowed,	4	
	or - not married?	5	
b)	And a few questions about you and your household. Including yourself, how many people live here regularly as members of this household? **CHECK INTERVIEWER MANUAL FOR DEFINITION OF HOUSEHOLD IF NECESSARY.**		
	WRITE IN:	1409-10	
901.	Now I'd like to ask for a few details about each person in your household. Starting with yourself, what was your age last birthday? **WORK DOWN COLUMNS OF GRID FOR EACH HOUSEHOLD MEMBER.**		

		Resp-ondent	2	3	4	5	6	7	8	9	10
a)	**Sex:**	11	15	20	25	30	35	40	45	50	55
	Male	1	1	1	1	1	1	1	1	1	1
	Female	2	2	2	2	2	2	2	2	2	2
		12-13	16-17	21-22	26-27	31-32	36-37	41-42	46-47	51-52	56-57
b)	**Age last birthday:**										

Question 900a (Marital status of respondent) is **precoded** and the interviewer will circle the appropriate code (1 to 5) in the box numbered 1408 in the right hand column. For question 900b (Number of people in the household) the interviewer will write the number (right justified or with leading zero) in the pair of boxes numbered 1409 - 10.

Question 901 is a typical **grid** for collecting information about members of the respondent's household. The layout is slightly more complex, but quite straightforward. Question 901a (Sex of respondent) is precoded and again the interviewer will circle the appropriate code (1 for Male, 2 for Female). For question 901b (Age of respondent last birthday) the interviewer will write in the age in the first pair of boxes on the line (numbered 12 - 13).

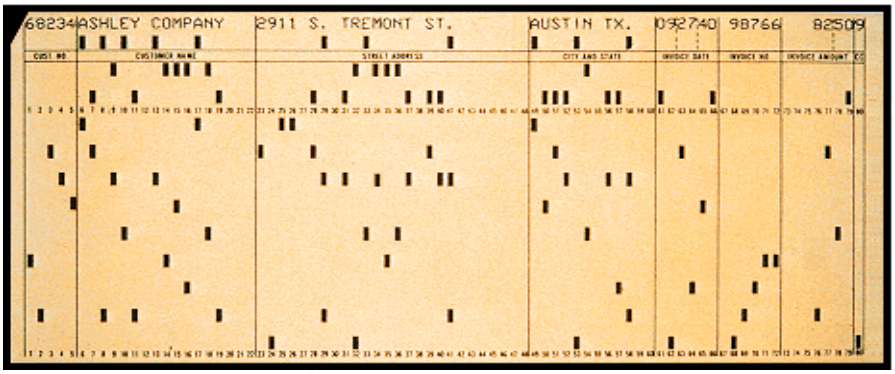
The grid continues (not shown) with questions on the relationship to the respondent of each additional person in the household and whether that person has legal responsibility for the accommodation. For now, this doesn't concern us, but if there are any other people in the household, the interviewer will, for each additional person, circle the code for sex and write in the age in the appropriate pair(s) of boxes, starting with the second person and coding sex in the box numbered 15 (at the top) and age in the boxes numbered 16-17.

How do the responses to these questions get into the computer?

The data from this survey were originally keyed in as fixed format 80-column lines (card images), using the data layout template printed in the right hand margins of the questionnaire. This format derives from the days before visual display units (VDUs) and keyboards, when all data (and also programs) had to be punched on 80-column Hollerith cards and fed into the computer via a card reader.

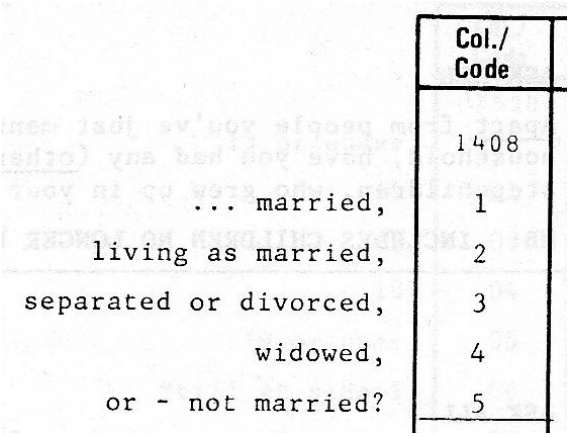
80-column Hollerith card

From Computer Desktop Encyclopedia
© 2000 The Computer Language Co. Inc.

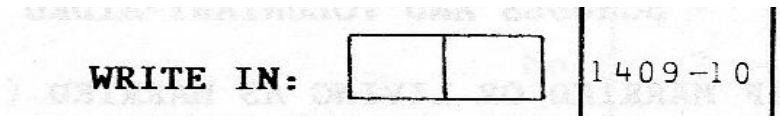


Have another look at the questionnaire.

In the top right hand corner, for question 900a (Marital status), you can see the number **1408** in smaller type in the margin and above the response codes 1 - 5.



…and further down, for question 900b a pair of boxes for the interviewer to write in the number of people in the household, next to **1409 -10** (in the margin).



These are special numbers used at the data preparation stage. In the 1970s and 1980s, before computer-assisted personal interviews (CAPI) data from completed questionnaires were usually punched on to 80-column Hollerith cards (sometimes 60- or 120- columns), which were then fed into the computer via a card reader.

The data could then be printed out on a lineprinter. Later developments enabled data to be keyed directly into the computer as fixed-format card-images and displayed on a computer screen, one

line for each card. In other words, the data for each response code were always punched in the same column(s) on the same card for every respondent.

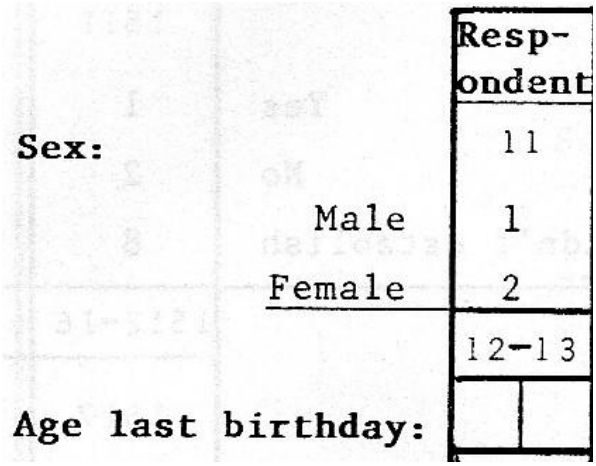
The 1989 British Social Attitudes survey consisted of two versions, A and B, with separate questionnaires and a separate self-completion questionnaire for each version. Each version carried a core of common questions plus a different a set of questions on key topics which were split between versions. Altogether there were 3,024 respondents, of whom 1,508 were administered version A and 1,516 version B. Each questionnaire takes up 23 lines of data and each line is numbered from 01 to 23.

On the computer, when you open the raw data file, each record is displayed as a line on the screen. There are 23 lines of data for each respondent, but, apart from the serial number and record number, the lines for the data from versions A or B[3] will be blank depending on which version the respondent completed. In each of the 23 records for each of the 3024 cases (i.e. every record in the entire data set) columns 1-5 contain a unique serial number and columns 6-7 the record number (from 01 to 23).

In this example, the data for questions 900 and 901 were punched on card 14. We know this because the data layout convention used on the printed questionnaire is indicated by the four digit number **1408** in small type at the top of the right hand margin (see above). In this convention the first pair of digits **14** indicate the card or line number (from 01 to 23) within each case and the second pair **08** the **field** (column(s)) within each card (from 01 to 80). Thus **1408** indicates that the data for Q.900a (Marital status) are to be punched on card 14 column 8: similarly data for Q.900b (Number in household) are on card 14 columns 9 and 10, as indicated by **1409-10** (two columns are needed for numbers greater than 9).

In SPSS terminology each card, card-image or data-line is now called a **record**.

Sex is coded as 1 or 2 and is to be punched as a single digit on record 14, column 11.
Age is to be punched as a two-digit number on record 14, columns 12 – 13.



[3] The code for which version was admiinistered is in column 8 of record 2 (viz, **208**)

Here are the actual data[4] for a single case from a respondent who completed version B of the questionnaire.

```
238290109012829191203        511201  2  0201150450201         1    61  00.6666
2382902220381511042210 311  1111121221  13113112112121132 13343132  031 3943122
23829034              1006  33312    2    42 090106122 222 2112114410390903
2382904     152311151          3
2382905
2382906
2382907                                 030114113331142122 113112     232344
2382908010407 11131210        1212112312     231211412111111
2382909
2382910
2382911412 4               2 12 4444  3 3
23829128  8  3831       1            2                               14122
2382913   1330342     1 1 1222  2212211211 226                           2222227
2382914102251 15911                                                  1
2382915 22      21  2        12                                  069010907329604
2382916    000010341 22222203     071030808327108    00061021111             1
23829172211    4        130611112 221221230459191250789                        3
2382918
2382919432221423223 1     43122312223   21    3 2122111  2223 1       333422224
2382920223535525211                                      32221122221232211511212
238292113413331223331                                       3422322222222222333
2382922                                     1      335333         2344444441311
238292323221  22322333323     225078921241064115        0102654424341008373215
```

This is a conversion of the original data from **Times New Roman** **variable-width** font to plain text in **Courier New** **fixed-width** font. The data are displayed in 80-column lines in which all 80 columns are vertically aligned.

Question: What are the sex and age of this respondent?

Try to answer this yourself before reading the next page.

[4] You can download the data bsa89.txt (3.9 mb) **from this site.**

The record number is in columns 6 – 7, highlighted in **pink**. Count down to find record 14 (highlighted in light turquoise) then count along to find **2** in column 11 (**sex**) and **51** in columns 12-13 (**age**).

```
238290109012829191203        511201  2  0201150450201         1    61  00.6666
2382902220381511042210 311  1111121221  13113112112121132 13343132  031 3943122
23829034              1006  33312    2    42 090106122 222 2112114410390903
2382904     152311151          3
2382905
2382906
2382907                                 030114113331142122 113112     232344
2382908010407 11131210        1212112312     231211412111111
2382909
2382910
2382911412 4               2 12 4444  3 3
23829128  8  3831       1            2                               14122
2382913   1330342     1 1 1222  2212211211 226                           2222227
2382914102251 15911                                                 1
2382915 22      21  2        12                                  069010907329604
2382916    000010341 22222203     071030808327108    00061021111             1
23829172211    4        130611112 2212212304591912250789                       3
2382919432221423223 1     43122312223   21    3 2122111  2223 1       333422224
2382920223535525211                                      32221122221232211511212
238292113413331223331                                       34223222222222222333
2382922                                    1      335333         2344444441311
2382923323221 22322333323     225078921241064115        01026544243410083732 15
```

This respondent is therefore a **woman** aged **51**. What is her **marital status**? How many people live in the accommodation? Which version of the questionnaire did she complete? Find the answers yourself!!

Now go to section **1.3: Reading raw data into SPSS**

Next step: **1.3.3.1 Preparing the ground**

[Back to Block 1 menu]