# Updating the Classics: a New Life for Old Data

Sharon Bolton

Data Publishing and Curation Manager

IASSIST & CARTO 2018: Once Upon a Data Point: Sustaining our Data Storytellers

Montreal, Canada, 31st May 2018

UK Data Service

UK • DATA ARCHIVE

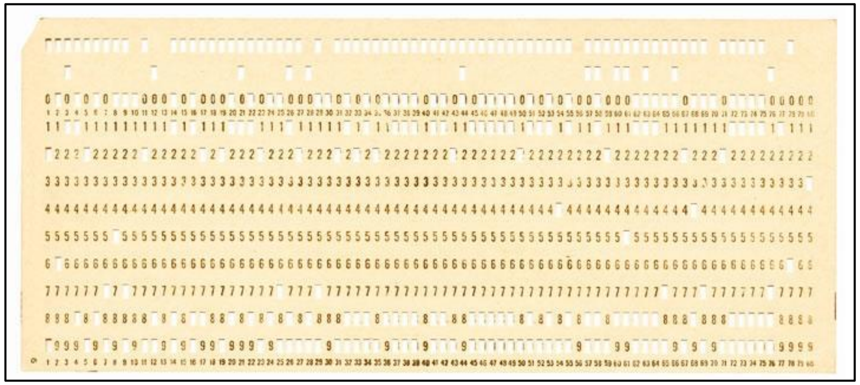
University of Essex

# Setting the scene

- 2016 Brexit referendum result:
  - What had changed since 1975? New interest in revisiting old research
- Worked with researcher to find the data needed
  - Public attitudes to UK foreign policy in late 1960s/early 1970s
  - Lots of literature but less data
  - National Opinion Polls (NOP) data had most potential
- Roper Center holds extensive collection, behind paywall
- UK Data Service holds many – some not used for a long time
  - Paper documentation scanned to PDF and available on web, but
  - Data in older ‘column binary’ format, difficult to use

# What is column binary data?

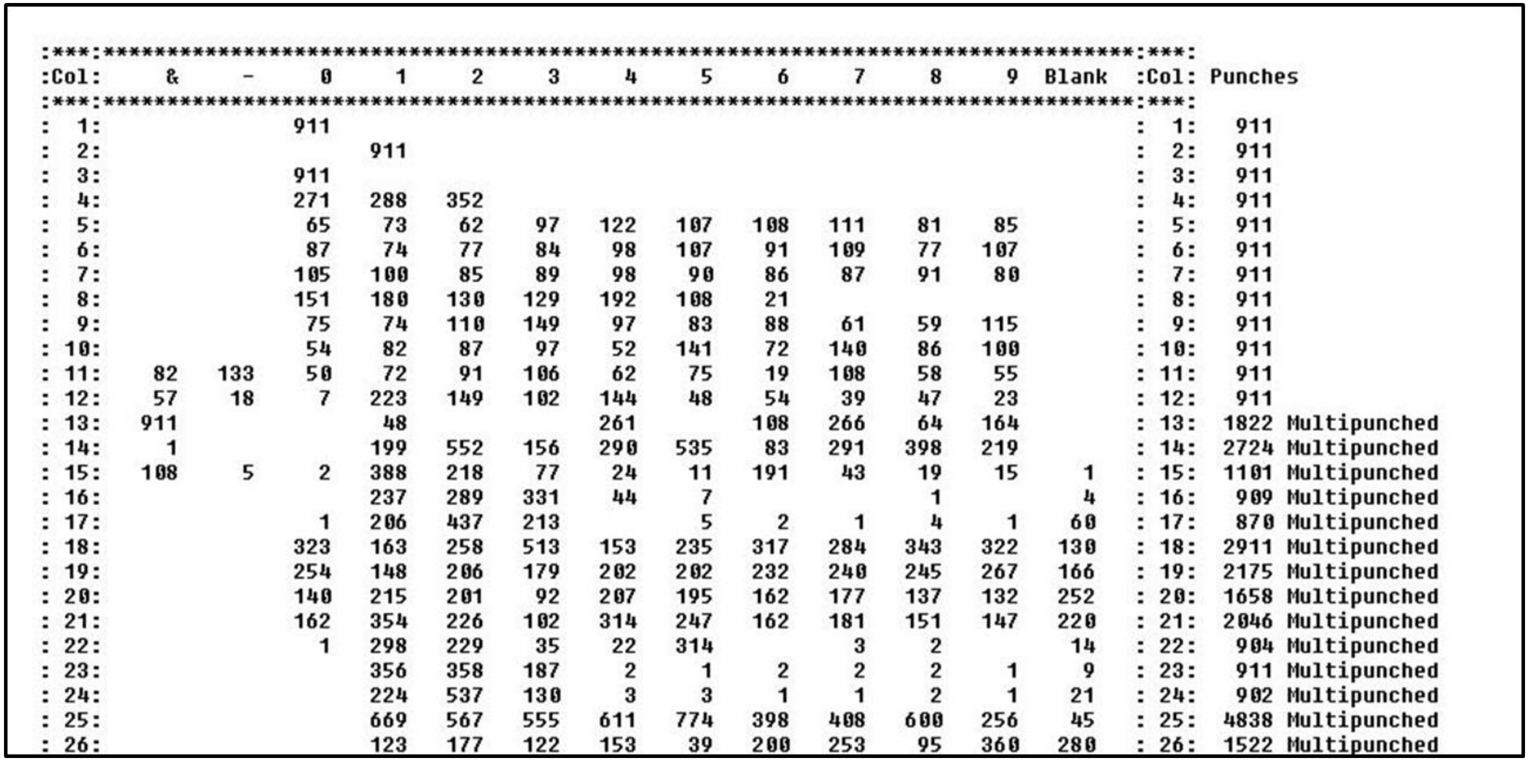
- Raw data once stored on computer punch cards - standard 80 columns of data occupies the 12 rows of each card. Data stored in this way are called column binary or multi-punch data, and allow more than one variable (in this case, survey question) to be stored in the same column  (adapted from Landman, 1996)



- Read by card reader machine that creates digital column binary data files
- Data can’t be read by current statistical packages without conversion

# Column binary layout

- 12 columns → 80 rows ↓
- Usually, & = value 11, - = 12, 0 = 10, 1-9 = 4-12, Blank=missing
- But not always – check the data!



```
:***:*******************************************************************************:***:
:Col:     &     -     0     1     2     3     4     5     6     7     8     9 Blank :Col: Punches
:***:*******************************************************************************:***:
:  1:               911                                                             :  1:   911
:  2:                     911                                                       :  2:   911
:  3:               911                                                             :  3:   911
:  4:               271   288   352                                                 :  4:   911
:  5:                65    73    62    97   122   107   108   111    81    85       :  5:   911
:  6:                87    74    77    84    98   107    91   109    77   107       :  6:   911
:  7:               105   100    85    89    98    90    86    87    91    80       :  7:   911
:  8:               151   180   130   129   192   108    21                         :  8:   911
:  9:                75    74   110   149    97    83    88    61    59   115       :  9:   911
: 10:                54    82    87    97    52   141    72   140    86   100       : 10:   911
: 11:    82   133    50    72    91   106    62    75    19   108    58    55       : 11:   911
: 12:    57    18     7   223   149   102   144    48    54    39    47    23       : 12:   911
: 13:   911                48                261         108   266    64   164       : 13:  1822 Multipunched
: 14:     1               199   552   156   290   535    83   291   398   219       : 14:  2724 Multipunched
: 15:   108     5     2   388   218    77    24    11   191    43    19    15     1 : 15:  1101 Multipunched
: 16:                     237   289   331    44     7                 1           4 : 16:   909 Multipunched
: 17:                 1   206   437   213           5     2     1     4     1    60 : 17:   870 Multipunched
: 18:               323   163   258   513   153   235   317   284   343   322   130 : 18:  2911 Multipunched
: 19:               254   148   206   179   202   202   232   240   245   267   166 : 19:  2175 Multipunched
: 20:               140   215   201    92   207   195   162   177   137   132   252 : 20:  1658 Multipunched
: 21:               162   354   226   102   314   247   162   181   151   147   220 : 21:  2046 Multipunched
: 22:                 1   298   229    35    22   314           3     2          14 : 22:   904 Multipunched
: 23:                     356   358   187     2     1     2     2     2     1     9 : 23:   911 Multipunched
: 24:                     224   537   130     3     3     1     1     2     1    21 : 24:   902 Multipunched
: 25:                     669   567   555   611   774   398   408   600   256    45 : 25:  4838 Multipunched
: 26:                     123   177   122   153    39   200   253    95   360   280 : 26:  1522 Multipunched
```

## How did we get here?

- UK Data Archive 50 years old – large collection over lifetime
- Collection management requires time and resources, funding constraints had meant other work prioritised
- ‘Data archaeologist’ used documents from long-finished Gallup Poll project to develop conversion script for column binary > SPSS (Landman, 1996)
- Had used method before but never on this scale: over 50 datasets in NOP series, researcher needed c.30 in a timely fashion

# Serendipity

- One-time extra funding agreed for collection management

- The mission – search and rescue
  - Find and convert the NOP series

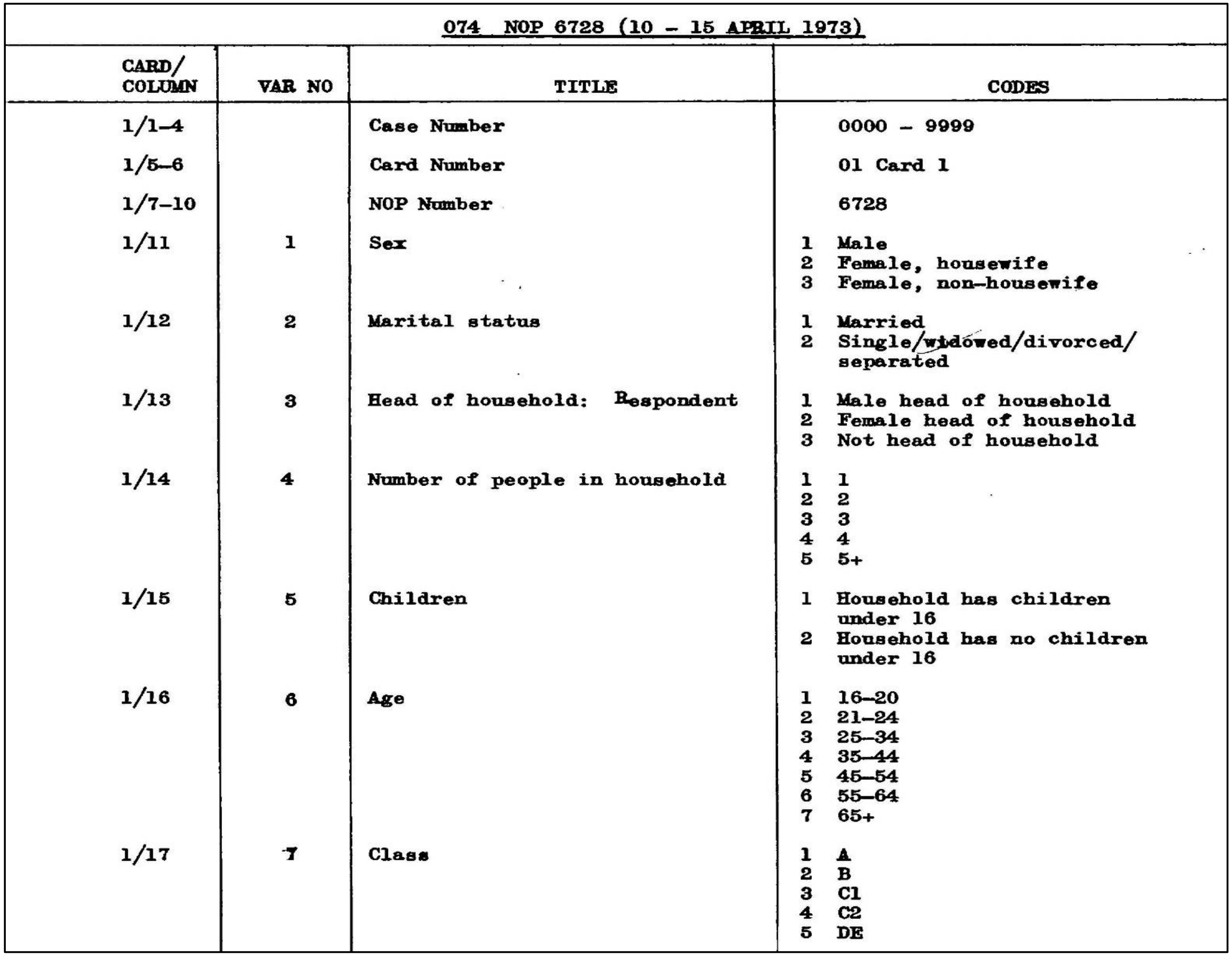
- The quest
  - Recruited a curator and trained him in the tools to do the job
  - Background in data analysis
  - Programming in SPSS, using and manipulating syntax, running scripts

- The goal
  - Make the NOP series data available and easy to use

# How do we do it?

- Make a map - tell the software where to look in the column binary data file and what to do with the information it finds
- Making the map depends on good metadata - if the map is correct you can find the treasure
- Did we have enough information to find our way? Write the script and see
- Documentation = curation notes, questionnaires, reports

# Metadata dream

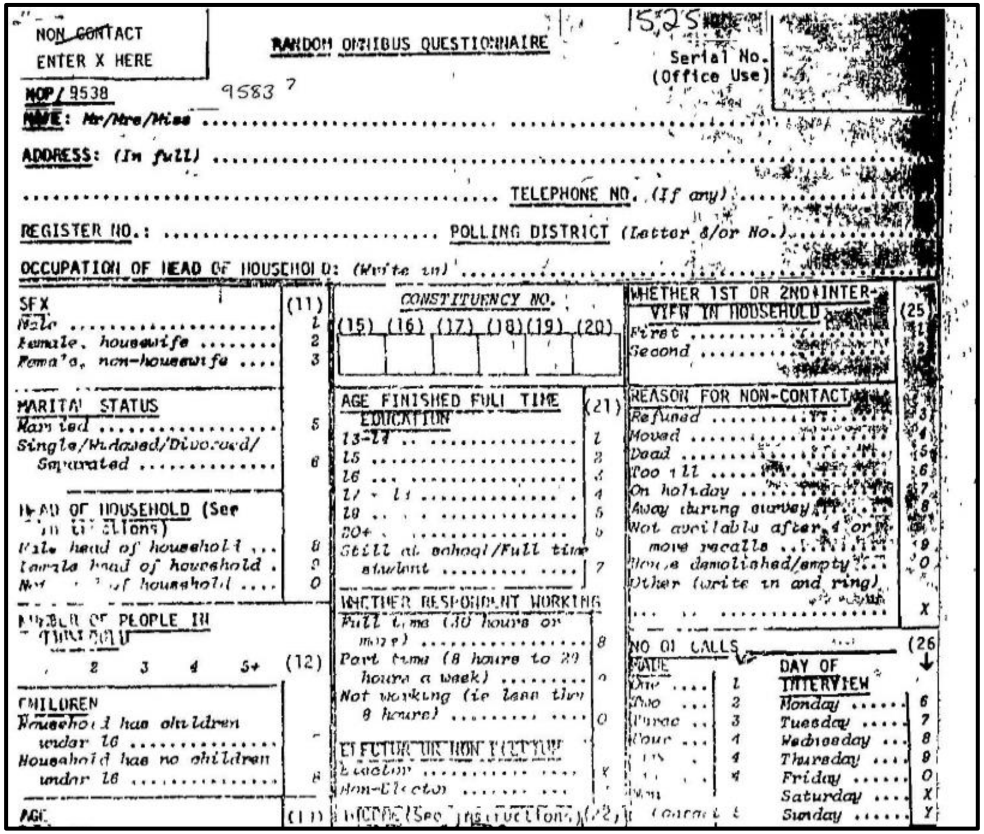


074 NOP 6728 (10 - 15 APRIL 1973)

| CARD/<br>COLUMN | VAR NO | TITLE | CODES |
|---|---|---|---|
| 1/1-4 | | Case Number | 0000 - 9999 |
| 1/5-6 | | Card Number | 01 Card 1 |
| 1/7-10 | | NOP Number | 6728 |
| 1/11 | 1 | Sex | 1 Male<br>2 Female, housewife<br>3 Female, non-housewife |
| 1/12 | 2 | Marital status | 1 Married<br>2 Single/widowed/divorced/ separated |
| 1/13 | 3 | Head of household: Respondent | 1 Male head of household<br>2 Female head of household<br>3 Not head of household |
| 1/14 | 4 | Number of people in household | 1 1<br>2 2<br>3 3<br>4 4<br>5 5+ |
| 1/15 | 5 | Children | 1 Household has children under 16<br>2 Household has no children under 16 |
| 1/16 | 6 | Age | 1 16-20<br>2 21-24<br>3 25-34<br>4 35-44<br>5 45-54<br>6 55-64<br>7 65+ |
| 1/17 | 7 | Class | 1 A<br>2 B<br>3 C1<br>4 C2<br>5 DE |

UK Data Service

# Metadata nightmare



NON CONTACT
ENTER X HERE

RANDOM OMNIBUS QUESTIONNAIRE

15̸25

Serial No.
(Office Use)

NOP/ 9538 9583 7

NAME: Mr/Mrs/Miss ..........................................

ADDRESS: (In full) ..........................................

.................................................. TELEPHONE NO. (If any) ..........

REGISTER NO.: .............................. POLLING DISTRICT (Letter &/or No.) ..........

OCCUPATION OF HEAD OF HOUSEHOLD: (Write in) ..........................................

| SEX | (11) |
|---|---|
| Male .......................... | 1 |
| Female, housewife .......... | 2 |
| Female, non-housewife .... | 3 |

| MARITAL STATUS | |
|---|---|
| Married ...................... | 5 |
| Single/Widowed/Divorced/ Separated .............. | 6 |

| HEAD OF HOUSEHOLD (See instructions) | |
|---|---|
| Male head of household ... | 8 |
| Female head of household . | 9 |
| Not head of household .... | 0 |

NUMBER OF PEOPLE IN HOUSEHOLD

| 1 | 2 | 3 | 4 | 5+ | (12) |
|---|---|---|---|---|---|

| CHILDREN | |
|---|---|
| Household has children under 16 .............. | 7 |
| Household has no children under 16 .............. | 8 |

AGE (13)

CONSTITUENCY NO.

| (15) | (16) | (17) | (18) | (19) | (20) |
|---|---|---|---|---|---|
| | | | | | |

| AGE FINISHED FULL TIME EDUCATION | (21) |
|---|---|
| 13-14 .. .............. | 1 |
| 15 .................... | 2 |
| 16 .................... | 3 |
| 17 - 18 ................ | 4 |
| 19 .................... | 5 |
| 20+ .................. | 6 |
| Still at school/Full time student .............. | 7 |

| WHETHER RESPONDENT WORKING | |
|---|---|
| Full time (30 hours or more) .............. | 8 |
| Part time (8 hours to 29 hours a week) ......... | 9 |
| Not working (ie less than 8 hours) ........ | 0 |

| ELECTOR OR NON ELECTOR | |
|---|---|
| Elector .............. | X |
| Non-Elector .............. | |

INCOME (See instructions) (22)

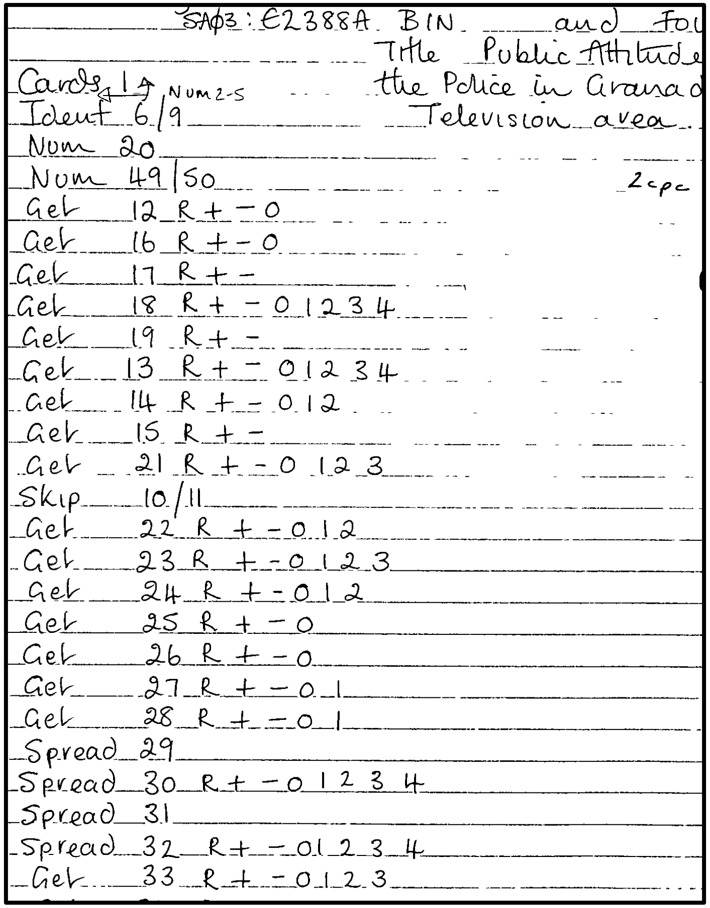
| WHETHER 1ST OR 2ND INTERVIEW IN HOUSEHOLD | (25) |
|---|---|
| First .................... | 1 |
| Second .................. | 2 |

| REASON FOR NON-CONTACT | |
|---|---|
| Refused .................. | 3 |
| Moved .................... | 4 |
| Dead ...................... | 5 |
| Too ill .................. | 6 |
| On holiday .............. | 7 |
| Away during survey ...... | 8 |
| Not available after 4 or more recalls ........ | 9 |
| House demolished/empty .. | 0 |
| Other (write in and ring) ......................... | X |

NO OF CALLS (26)

| MADE | |
|---|---|
| One .... | 1 |
| Two .... | 2 |
| Three .. | 3 |
| Four ... | 4 |
| Five ... | 4 |
| ... | 4 |
| Wen (contact) | 5 |

| DAY OF INTERVIEW | |
|---|---|
| Monday ...... | 6 |
| Tuesday ..... | 7 |
| Wednesday ... | 8 |
| Thursday .... | 9 |
| Friday ...... | 0 |
| Saturday .... | X |
| Sunday ...... | Y |

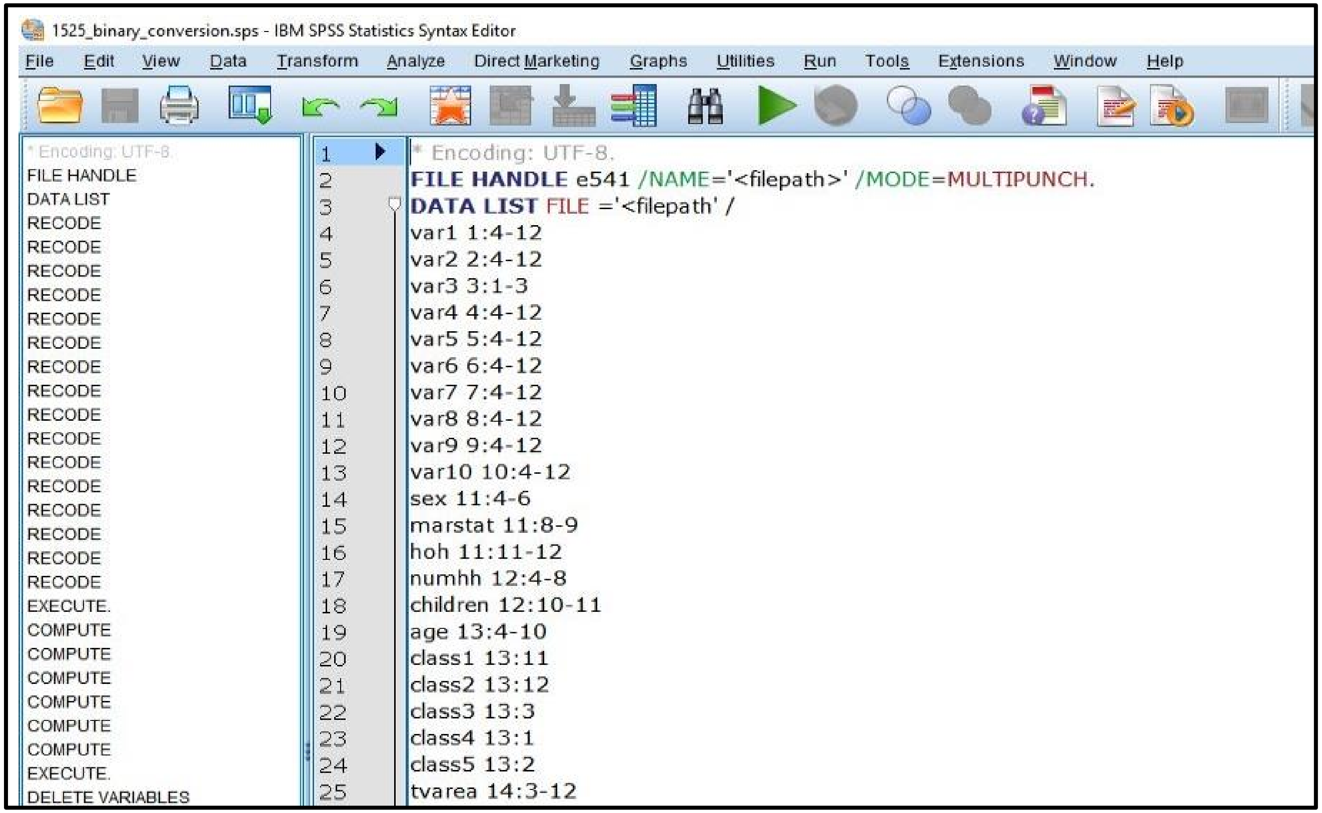


SAϕ3 : E2388A BIN and Fo
Title Public Attitude
the Police in Granad
Television area

Cards 1 Num 2-5
Ident 6/9
Num 20
Num 49/50 2cpc
Get 12 R + − 0
Get 16 R + − 0
Get 17 R + −
Get 18 R + − 0 1 2 3 4
Get 19 R + −
Get 13 R + − 0 1 2 3 4
Get 14 R + − 0 1 2
Get 15 R + −
Get 21 R + − 0 1 2 3
Skip 10/11
Get 22 R + − 0 1 2
Get 23 R + − 0 1 2 3
Get 24 R + − 0 1 2
Get 25 R + − 0
Get 26 R + − 0
Get 27 R + − 0 1
Get 28 R + − 0 1
Spread 29
Spread 30 R + − 0 1 2 3 4
Spread 31
Spread 32 R + − 0 1 2 3 4
Get 33 R + − 0 1 2 3

# Drawing the map

- Write and format the script – no shortcuts
- Trial and error: run the script, check the results against the documentation – multipunch columns can cause errors
- Amend the script, run it again until data correct

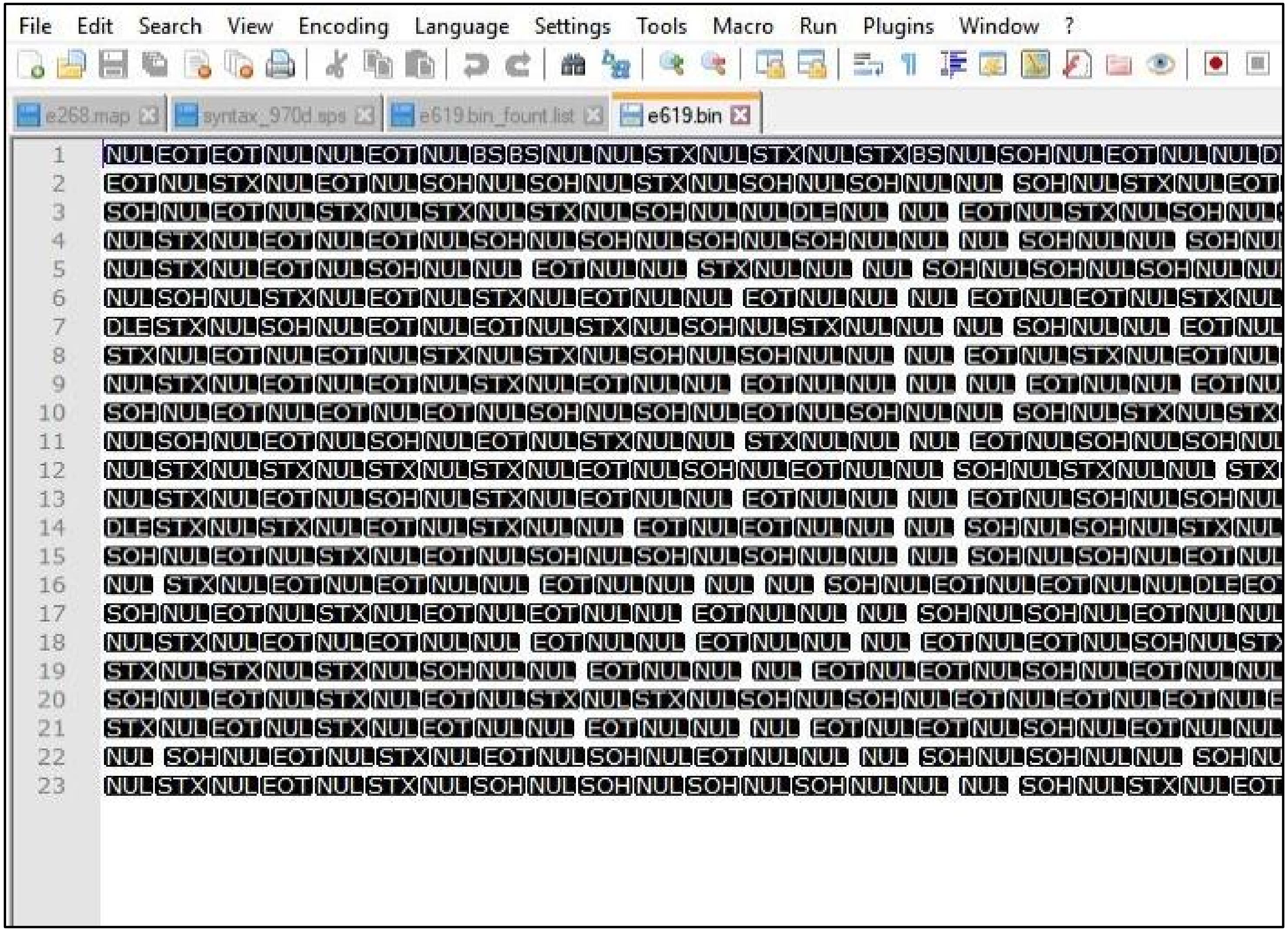


```
* Encoding: UTF-8.
FILE HANDLE e541 /NAME='<filepath>' /MODE=MULTIPUNCH.
DATA LIST FILE ='<filepath' /
var1 1:4-12
var2 2:4-12
var3 3:1-3
var4 4:4-12
var5 5:4-12
var6 6:4-12
var7 7:4-12
var8 8:4-12
var9 9:4-12
var10 10:4-12
sex 11:4-6
marstat 11:8-9
hoh 11:11-12
numhh 12:4-8
children 12:10-11
age 13:4-10
class1 13:11
class2 13:12
class3 13:3
class4 13:1
class5 13:2
tvarea 14:3-12
```

# Completing the job

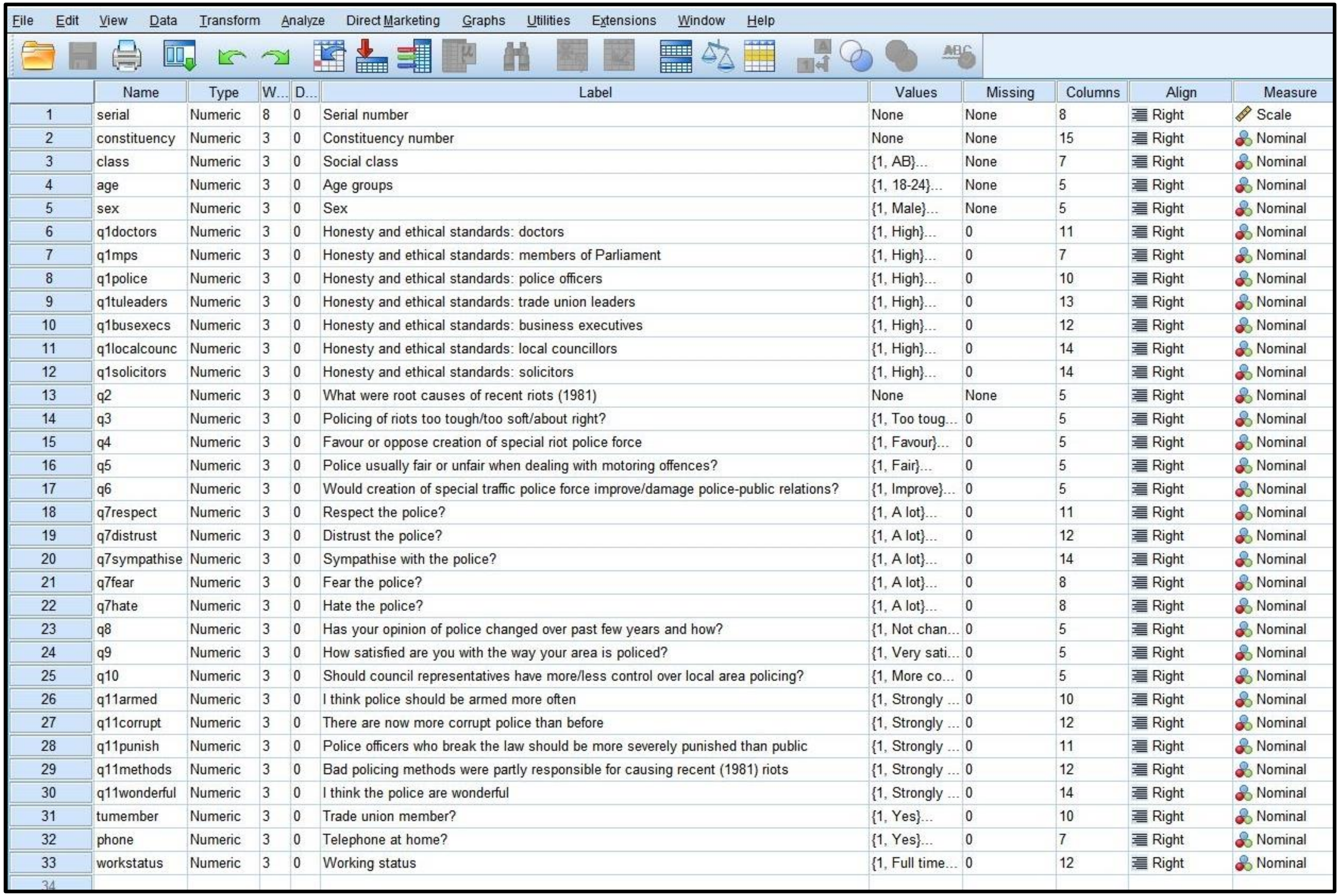
- Clean and label the data – recoding multi-punch variables, string characters to numeric, add metadata (variable and value labels)
- Enhance usability - apply robust UK Data Service curation standards
- Create preservation format (ASCII) and preservation metadata
- Create current standard dissemination formats – SPSS, Stata, tab-delimited text
- Upgrade scanned documentation – optical character recognition (OCR), PDF/A where possible
- Augment catalogue metadata

# Before …



```
NUL EOT EOT NUL NUL EOT NUL BS BS NUL NUL STX NUL STX NUL STX BS NUL SOH NUL EOT NUL NUL D
EOT NUL STX NUL EOT NUL SOH NUL SOH NUL STX NUL SOH NUL SOH NUL NUL  SOH NUL STX NUL EOT
SOH NUL EOT NUL STX NUL STX NUL STX NUL SOH NUL NUL DLE NUL  NUL  EOT NUL STX NUL SOH NUL
NUL STX NUL EOT NUL EOT NUL SOH NUL SOH NUL SOH NUL SOH NUL NUL  NUL  SOH NUL NUL  SOH NU
NUL STX NUL EOT NUL SOH NUL NUL  EOT NUL NUL  STX NUL NUL  NUL  SOH NUL SOH NUL SOH NUL NU
NUL SOH NUL STX NUL EOT NUL STX NUL EOT NUL NUL  EOT NUL NUL  NUL  EOT NUL EOT NUL STX NUL
DLE STX NUL SOH NUL EOT NUL EOT NUL STX NUL SOH NUL STX NUL NUL  NUL  SOH NUL NUL  EOT NUL
STX NUL EOT NUL EOT NUL STX NUL STX NUL SOH NUL SOH NUL NUL  NUL  EOT NUL STX NUL EOT NUL
NUL STX NUL EOT NUL EOT NUL STX NUL EOT NUL NUL  EOT NUL NUL  NUL  NUL  EOT NUL NUL  EOT NU
SOH NUL EOT NUL EOT NUL EOT NUL SOH NUL SOH NUL EOT NUL SOH NUL NUL  SOH NUL STX NUL STX
NUL SOH NUL EOT NUL SOH NUL EOT NUL STX NUL NUL  STX NUL NUL  NUL  EOT NUL SOH NUL SOH NUL
NUL STX NUL STX NUL STX NUL STX NUL EOT NUL SOH NUL EOT NUL NUL  SOH NUL STX NUL NUL  STX
NUL STX NUL EOT NUL SOH NUL STX NUL EOT NUL NUL  EOT NUL NUL  NUL  EOT NUL SOH NUL SOH NUL
DLE STX NUL STX NUL EOT NUL STX NUL NUL  EOT NUL EOT NUL NUL  NUL  SOH NUL SOH NUL STX NUL
SOH NUL EOT NUL STX NUL EOT NUL SOH NUL SOH NUL SOH NUL NUL  NUL  SOH NUL SOH NUL EOT NUL
NUL  STX NUL EOT NUL EOT NUL NUL  EOT NUL NUL  NUL  NUL  SOH NUL EOT NUL EOT NUL NUL DLE EOT
SOH NUL EOT NUL STX NUL EOT NUL EOT NUL NUL  EOT NUL NUL  NUL  SOH NUL SOH NUL EOT NUL NUL
NUL STX NUL EOT NUL EOT NUL NUL  EOT NUL NUL  EOT NUL NUL  NUL  EOT NUL EOT NUL SOH NUL STX
STX NUL STX NUL STX NUL SOH NUL NUL  EOT NUL NUL  NUL  EOT NUL EOT NUL SOH NUL EOT NUL NUL
SOH NUL EOT NUL STX NUL EOT NUL STX NUL STX NUL SOH NUL SOH NUL EOT NUL EOT NUL EOT NUL E
STX NUL EOT NUL STX NUL EOT NUL NUL  EOT NUL NUL  NUL  EOT NUL EOT NUL SOH NUL EOT NUL NUL
NUL  SOH NUL EOT NUL STX NUL EOT NUL SOH NUL EOT NUL NUL  NUL  SOH NUL SOH NUL NUL  SOH NU
NUL STX NUL EOT NUL STX NUL SOH NUL SOH NUL SOH NUL SOH NUL NUL  NUL  SOH NUL STX NUL EOT
```

# After …



| | Name | Type | W... | D... | Label | Values | Missing | Columns | Align | Measure |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | serial | Numeric | 8 | 0 | Serial number | None | None | 8 | Right | Scale |
| 2 | constituency | Numeric | 3 | 0 | Constituency number | None | None | 15 | Right | Nominal |
| 3 | class | Numeric | 3 | 0 | Social class | {1, AB}... | None | 7 | Right | Nominal |
| 4 | age | Numeric | 3 | 0 | Age groups | {1, 18-24}... | None | 5 | Right | Nominal |
| 5 | sex | Numeric | 3 | 0 | Sex | {1, Male}... | None | 5 | Right | Nominal |
| 6 | q1doctors | Numeric | 3 | 0 | Honesty and ethical standards: doctors | {1, High}... | 0 | 11 | Right | Nominal |
| 7 | q1mps | Numeric | 3 | 0 | Honesty and ethical standards: members of Parliament | {1, High}... | 0 | 7 | Right | Nominal |
| 8 | q1police | Numeric | 3 | 0 | Honesty and ethical standards: police officers | {1, High}... | 0 | 10 | Right | Nominal |
| 9 | q1tuleaders | Numeric | 3 | 0 | Honesty and ethical standards: trade union leaders | {1, High}... | 0 | 13 | Right | Nominal |
| 10 | q1busexecs | Numeric | 3 | 0 | Honesty and ethical standards: business executives | {1, High}... | 0 | 12 | Right | Nominal |
| 11 | q1localcounc | Numeric | 3 | 0 | Honesty and ethical standards: local councillors | {1, High}... | 0 | 14 | Right | Nominal |
| 12 | q1solicitors | Numeric | 3 | 0 | Honesty and ethical standards: solicitors | {1, High}... | 0 | 14 | Right | Nominal |
| 13 | q2 | Numeric | 3 | 0 | What were root causes of recent riots (1981) | None | None | 5 | Right | Nominal |
| 14 | q3 | Numeric | 3 | 0 | Policing of riots too tough/too soft/about right? | {1, Too toug... | 0 | 5 | Right | Nominal |
| 15 | q4 | Numeric | 3 | 0 | Favour or oppose creation of special riot police force | {1, Favour}... | 0 | 5 | Right | Nominal |
| 16 | q5 | Numeric | 3 | 0 | Police usually fair or unfair when dealing with motoring offences? | {1, Fair}... | 0 | 5 | Right | Nominal |
| 17 | q6 | Numeric | 3 | 0 | Would creation of special traffic police force improve/damage police-public relations? | {1, Improve}... | 0 | 5 | Right | Nominal |
| 18 | q7respect | Numeric | 3 | 0 | Respect the police? | {1, A lot}... | 0 | 11 | Right | Nominal |
| 19 | q7distrust | Numeric | 3 | 0 | Distrust the police? | {1, A lot}... | 0 | 12 | Right | Nominal |
| 20 | q7sympathise | Numeric | 3 | 0 | Sympathise with the police? | {1, A lot}... | 0 | 14 | Right | Nominal |
| 21 | q7fear | Numeric | 3 | 0 | Fear the police? | {1, A lot}... | 0 | 8 | Right | Nominal |
| 22 | q7hate | Numeric | 3 | 0 | Hate the police? | {1, A lot}... | 0 | 8 | Right | Nominal |
| 23 | q8 | Numeric | 3 | 0 | Has your opinion of police changed over past few years and how? | {1, Not chan... | 0 | 5 | Right | Nominal |
| 24 | q9 | Numeric | 3 | 0 | How satisfied are you with the way your area is policed? | {1, Very sati... | 0 | 5 | Right | Nominal |
| 25 | q10 | Numeric | 3 | 0 | Should council representatives have more/less control over local area policing? | {1, More co... | 0 | 5 | Right | Nominal |
| 26 | q11armed | Numeric | 3 | 0 | I think police should be armed more often | {1, Strongly ... | 0 | 10 | Right | Nominal |
| 27 | q11corrupt | Numeric | 3 | 0 | There are now more corrupt police than before | {1, Strongly ... | 0 | 12 | Right | Nominal |
| 28 | q11punish | Numeric | 3 | 0 | Police officers who break the law should be more severely punished than public | {1, Strongly ... | 0 | 11 | Right | Nominal |
| 29 | q11methods | Numeric | 3 | 0 | Bad policing methods were partly responsible for causing recent (1981) riots | {1, Strongly ... | 0 | 12 | Right | Nominal |
| 30 | q11wonderful | Numeric | 3 | 0 | I think the police are wonderful | {1, Strongly ... | 0 | 14 | Right | Nominal |
| 31 | tumember | Numeric | 3 | 0 | Trade union member? | {1, Yes}... | 0 | 10 | Right | Nominal |
| 32 | phone | Numeric | 3 | 0 | Telephone at home? | {1, Yes}... | 0 | 7 | Right | Nominal |
| 33 | workstatus | Numeric | 3 | 0 | Working status | {1, Full time... | 0 | 12 | Right | Nominal |
| 34 | | | | | | | | | | |

# End of Part One

- Special funding finished, back to conversion where and when we can

- Celebrate achievements so far
  - Impact! Researcher soon to publish book:

    Clements, B. *Public Opinion towards Foreign and Defence Policy in Britain, 1945-2017* (forthcoming, Routledge)

  - Not just NOP series, but >100 column binary datasets upgraded to current formats

  - Proven conversion methodology – scripts and algorithms work

  - Trained curator in useful data science skills

# Looking forward to Part Two

- Remaining column binary datasets to convert
- Modify scripts to work in other software (R, SAS, others?)
- Make scripts available to others, user guide, GitHub?
- These data born digital to UK Data Archive, need machine to read hard copy cards
- Renewed interest in data rescue (Research Data Alliance (RDA) Data Rescue IG)
- New funding opportunities for collection management?

# Questions

Sharon Bolton

sharonb@essex.ac.uk