Excel for Archivists Workshop

Data Improvement and Data Migration

Part 1 Data Improvement
(Units G – K)

HANDBOOK

Gillian Sheldrick
2022

Version 2.0

[blank page]

Excel for Archivists: Data Improvement and Data Migration
HANDBOOK CONTENTS

PREFACE AND ACKNOWLEDGEMENTS

DATA IMPROVEMENT THEME

G		**DATA IMPROVEMENT**	
	G1	Data improvement: an introduction	6
	G2	Subdividing columns of data	9
	G3	Removing duplicate records	13
H		**FORMULAS: TECHNIQUES AND REFERENCE**	
	H1	Techniques for working with formulas	18
	H2	Relative and absolute cell references	24
	H3	Useful functions and symbols	27
J		**TRANSFORMING DATA WITH FUNCTIONS AND FORMULAS**	
	J1	Conditional formulas	30
	J2	Combining formulas	37
	J3	Transforming data using a look-up list	44
K		**RESOLVING COMMON PROBLEMS**	
	K1	Understanding leading apostrophes and other special characters	48
	K2	Techniques for removing leading apostrophes	52

DATA MIGRATION THEME

M		**DATA MIGRATION: WHAT YOU NEED TO KNOW**	
	M1	Introduction to data migration	3
	M2	Data compatibility and data structure	5
	M3	Representing the archival hierarchy	8
	M4	Planning a data migration project	15
	M5	Data Mapping	18
N		**DATA MIGRATION TECHNIQUES**	
	N1	Importing into Excel using character separated values	31
	N2	Exporting from Excel using character separated values	36
	N3	Introducing XML and EAD	40
	N4	Exporting from Excel to XML	45
P		**MIGRATION AND SPECIALIST SYSTEMS**	
	P1	Migration between Excel and specialist systems: introductory	51
	P2	Microsoft Access	52
	P3	Axiell Calm	55
	P4	Axiell Adlib	59
	P5	The Archives Hub	62

PREFACE

This handbook forms part of the teaching material for a series of one-day training workshops aimed at archivists using Excel for transforming data or preparing it for migration into other systems. The workshops follow on from the workshop *Excel for Archivists: Essential Techniques and Knowledge*

The Workshops are specifically for archivists, are led by an archivist, and the examples and exercises relate to a series of specially prepared Excel workbooks based on genuine archive lists and catalogues

USING THIS HANDBOOK

Each unit relates to a single topic or a group of closely related topics. Most units consist of a series of practical worked examples, using a specially prepared Excel workbook. Others contain a set of notes on the theory. Most also include some exercises for practice. Many of the accompanying Excel workbooks contain several copies of the same worksheet, to allow for more practice without having to make a copy.

EXCEL VERSIONS

This handbook and the associated Excel workbooks are suitable for use with any version of Excel 2007 and later, although users of earlier versions will still find many of the concepts applicable. Note however that the illustrations in this Handbook mainly show Excel 2007. The differences between these and later versions are minor: the most obvious are the 'File' tab in version 2010 and later, which replaces the earlier 'Office' button, and the increased use of icons in the later versions.

ABOUT THE AUTHOR

Gillian Sheldrick has worked as an archivist for over forty years, and has a particular interest in cataloguing, as well as being a long-standing Excel enthusiast. She was Senior Archivist at Hertfordshire Archives and Local Studies, where she was responsible for the 'behind the scenes' aspects of the service, including the cataloguing programme, and then Cataloguing Manager for English Heritage where she was responsible for the new archive cataloguing system (built in-house) and for making the catalogues available on line. She now works as a freelance cataloguer and Excel trainer. In 2019 she qualified as one of the first Fellows of the Archives and Records Association (FARA).

Her interest in using IT to assist with cataloguing large collections dates back to the 1980s when the Archives service was considering buying a computer (by no means an obvious investment then). She was asked to investigate whether any use might be found for it and began experimenting with using simple databases. Since then she has moved on to Excel, which is much more powerful than any of those early databases. Excel for Archivists workshops began in 2014; to date (August 2022) over 50 Workshops and 20 online sessions have been delivered, with a total attendance of nearly 600.

ACKNOWLEDGEMENTS
The Excel workbooks used in the Workshop, to which this Handbook relates, are based on real lists and catalogues of archive collections (including early drafts), but have been adapted for use as teaching material. They should not be used as genuine finding aids.

Workshop participants are welcome to take copies of the workbooks for personal study and practice but they (or the Handbook) should not be distributed further.

I am grateful to the following for permission to use parts of lists originally compiled on their behalf:
Gloucestershire Archives
Churchill Archives Centre, Cambridge
The Centre for Scientific Archives
Wiltshire and Swindon History Centre

I would also like to thank friends, family and colleagues who suggested the idea of the workshop and who have helped with trialling and testing some of the material. I am also very grateful to workshop participants and other colleagues for their encouragement and for making suggestions for amendments and improvements.

I am particularly grateful to Jane Stevenson (The Archives Hub) and Dave Forster (Axiell Adlib) for their advice on several aspects of data migration; any errors are, of course, my own responsibility.

CONTACT DETAILS
For advice on using Excel, or to join the Excel for Archivists mailing list (mailings normally sent two or three times a year) contact me on SheldrickG@gmail.com

Gillian Sheldrick BA, BSc, DAS, DMS, FARA
(revised August 2022)

UNIT G1
DATA IMPROVEMENT: AN INTRODUCTION

Purpose: This unit provides an introduction to the theme of data improvement, including cross references to other relevant Units

1 Data Improvement or Data Cleaning

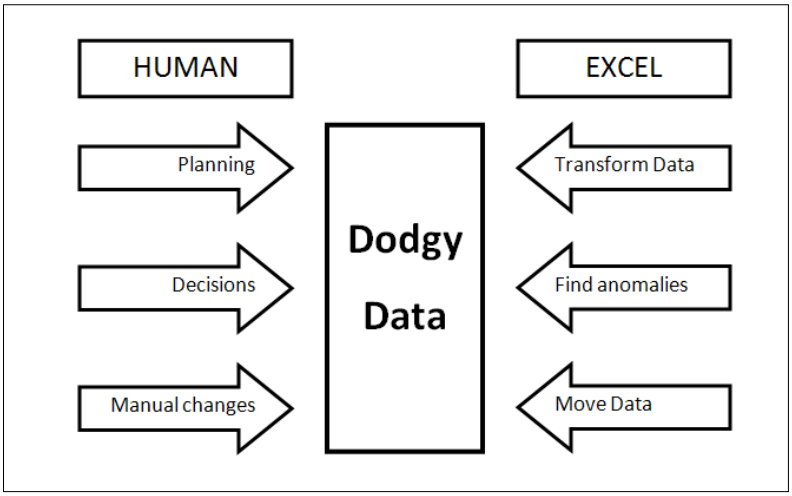
- Anything which improves existing data by transforming it in some way
- Minor changes (examples: correcting mis-spelling or altering the column order)
- Major changes (example: combining and standardizing data from multiple sources created to different data standards)

2 Types of change

- Correcting errors (example: spelling errors or inaccurate dates)
- Making the data easier to understand (example: expanding abbreviations)
- Making the data easier to use (example: sorting it into a logical order)
- Making the data internally consistent (example: standardising descriptive terms)
- Making the data consistent with an external standard (example: dividing it into appropriate columns before a data migration)

3 Excel can't do all the work

- Don't forget the human element



4 Useful techniques

Technique	Notes	Excel For Archivists UNIT reference
Spell check	See the Excel Review Ribbon. The Excel spell check is not as powerful as that in Word	Not covered
Filtering	Does not alter the data, but invaluable for identifying data anomalies	UNIT 2.1
Find and Replace		UNIT 2.4
Sorting rows	Including sorting subsets of data and sorting using multiple criteria	UNIT 2.3
Cut copy and paste	To rearrange data	UNIT 2.5
Subdividing columns of data		UNIT G2
Remove duplicate records		UNIT G3
Functions which provide information	Functions such as LEN (providing the number of characters), SEARCH (finds a specific character or text string) or COUNTA (the number of non-blank cells)	UNIT H3
Functions which change all text in the same way	Functions such as LEFT, RIGHT, CONCATENATE, UPPER apply a fixed rule to transform the data from one or more cells.	UNIT E1, E2, 3.4
Functions which act on text selectively	Functions which change data in different ways depending on its content. For example, VLOOKUP (changes a value to a new one specified in a lookup list); TRIM (removes excess spaces if any) or CLEAN (removes any nonprinting characters)	UNIT H3 and J3
Conditional formulas	Formulas introduced with =IF, which operate only if the conditions specified are met	UNIT J1
Combined formulas	Functions and formulas can act successively on the results of other functions and formulas. By applying the correct syntax they can also be combined into a single complex formula. If you can think of it, formulas can probably do it!	UNIT J2
Move the data out of Excel	It may sometimes be appropriate or convenient to move data out of Excel for editing (eg into Word, or a text editor).	For a case study, see UNITS K1 and K2 (leading apostrophes). See also the units relating to Data Migration (UNITS M, N and P)

For a case study in combining techniques (rearranging columns, applying filters, and using functions and formulas) see Excel for Archivists UNIT B3, *Creating a printed archive catalogue using Excel and Word*.

5 Planning a data cleaning project

Decide what you want to achieve

Be creative: There are usually different ways to achieve the same end

Most data cleaning operations combine multiple techniques: identify the steps and consider the order they need to be carried out

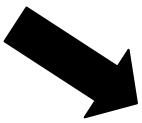
Consider the time involved: does the end result need to be perfect or merely adequate?

If only a few changes are needed, it may be quicker to retype the data rather than using a complicated Excel technique

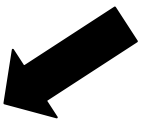
Allow plenty of time : don't rush

Take a back up copy before you start (and possibly after each step)

With large amounts of data, practise on a small sample first



REMEMBER: ALWAYS TAKE A BACK UP COPY BEFORE YOU START ALTERING DATA!!!!!



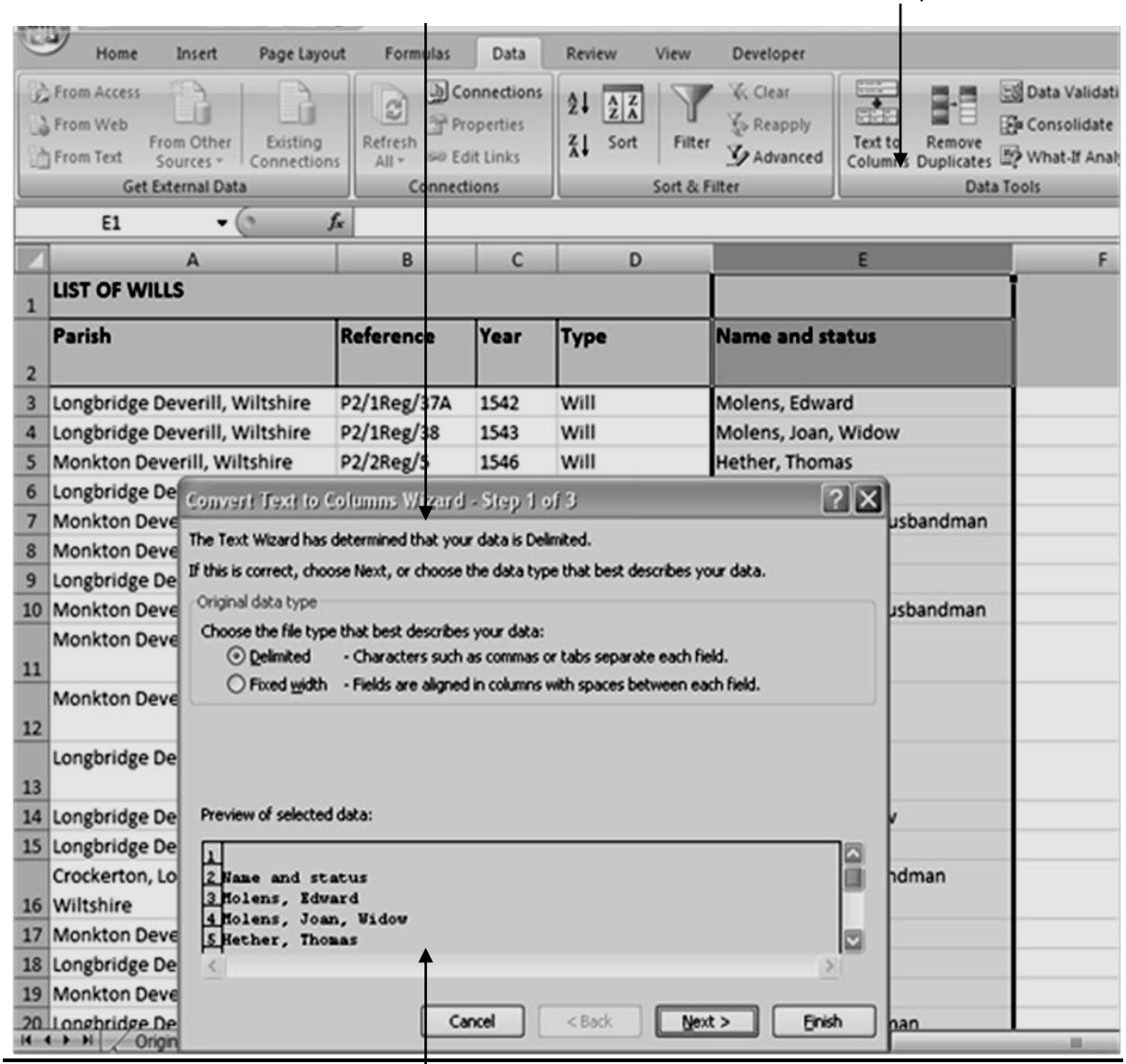
<< END OF THIS UNIT >>

UNIT G2
SUBDIVIDING COLUMNS OF DATA

Purpose: To separate data separated by commas, spaces or other characters into multiple columns

Convert Text to Columns Wizard

Text to columns
in Data Tools Group



Data Preview

A CONVERT TEXT TO COLUMNS WIZARD

➔ *Open the* **BLUEBERRY** *Workbook*

1 Open the BLUEBERRY Workbook
View the first worksheet entitled 'Original Data', which contains a list of probate records

- *The sheet contains the data before any transformations are carried out*
- *It also serves as a backup copy*

2 In the worksheet entitled 'Original Data' observe that 'Name and status' column (column C) contains data in the format *Marchant, Robert, Husbandman.* We wish to divide this between three columns (first name, last name and status).

3 In the BLUEBERRY Workbook, open the worksheet 'Copy 1 Names'.

- *The 'Name and Status' column has been moved to the right (column E). This is not essential, but makes the subsequent steps easier*

4 Go to the Data Ribbon and locate the Data Tools Group
Select column E (Name and status)
Click on Data Ribbon > Data Tools Group > Text to Columns

- *Convert Text to Columns Wizard Step 1 opens*
- *At the foot of the Wizard the data appears in a preview pane*

5 In the Convert Text to Columns Wizard Step 1, ensure the 'Delimited' button is selected
Then choose 'Next'

- *Convert Text to Columns Wizard Step 2 opens*

6 In Convert Text to Columns Wizard Step 2, select the 'comma' box

- *The preview pane changes to show the data split into separate columns*

7 In Convert Text to Columns Wizard Step 2, Step 2, choose 'Next'

- *Convert Text to Columns Wizard Step 3 opens*
- *In the preview pane, each new column can be selected in turn and the Data Type format can be changed if required*

8 In Convert Text to Columns Wizard Step 3, 'destination' box, type F1 (replacing the default value)
Then click 'Finish'
[If a message appears "Do you want to replace the contents of selected cells?" choose OK]

- *Convert Text to Columns Wizard closes*
- *The data from 'Name and status' appears in columns F, G, H and I, with separate columns replacing the commas*

9 Now examine the new data for consistency, for example

- *In Row 61 Smith, John, senior, Potter is now in four columns: elsewhere the occupation (Potter) is in column H, but in this case it is in column I. This might cause problems if the object of the exercise was to provide a list of all occupations*
- *In Row 28 Adlam alias Clevelodd, Magdalen, Widow the surname columns contains the term 'Adlam alias Clevelodd'; this might cause problems if the two surnames Adlam and Clevelodd were expected to be listed separately*
- *In row 111 Long, Manasseth (Manasses), Yeoman the first name appears as 'Manasseth (Manasses)' which might not be wanted.*

10 As with any automated data cleaning, it is essential to check that the result is as expected. If anomalies are found one of the following steps might be taken:

- *Amend the new data manually*
- *Change the automated routine*
- *Amend the data before carrying out the routine*

11 In this case, depending on the requirements for the new data, one or more of the following steps might be appropriate:

- *Apply filters to view only the few non-blank rows in column I, and then move the stray terms manually to column H (note that filters are available on the Data Ribbon as well as on the Home Ribbon)*
- *Using additional or alternative characters to divide the data (for example, dividing the data at spaces or brackets as well as commas)*
- *Use search and replace BEFORE applying the Text to columns routine, in order to remove the commas before 'senior' and 'junior' (for example, changing 'King, Thomas, senior, Yeoman' to 'King, Thomas senior, Yeoman')*

B CONVERT TEXT TO COLUMNS : NOTES

12

- Convert Text to Columns removes formatting and data types (number formats) so you may need to re-format after using the Wizard, or format the destination columns before applying text to Columns

- This can be a useful feature if you wish to remove formatting from a column

13 Convert Text to Columns can be a useful tool for analysing data divided with a consistent character. For example, splitting lengthy reference numbers in the format A/1/343/2/1/4 into separate columns can provide a useful visual indicator of levels of description (the lowest level references have data in every column, while higher level references such as A/1/343 are blank in the three right hand columns).

SUBDIVIDING COLUMNS OF DATA
FURTHER PRACTICE

➔ *Open the BLUEBERRY Workbook* which contains a list of probate records

Exercise 1

In the BLUEBERRY Workbook, open the worksheet entitled 'Copy 2 Reference', which contains another copy of the data, with the reference number moved to the right. Divide the reference number into separate columns, separated at the slash (/)

- ***HINT:*** *in Step 2 of the Wizard, select the type 'other' and type* ***/***

Exercise 2

In the BLUEBERRY Workbook, use the worksheet entitled Copy 3, Copy 4 or Copy 5. Divide the Parish into separate columns

(a) separated at commas
(b) separated at spaces

Exercise 3

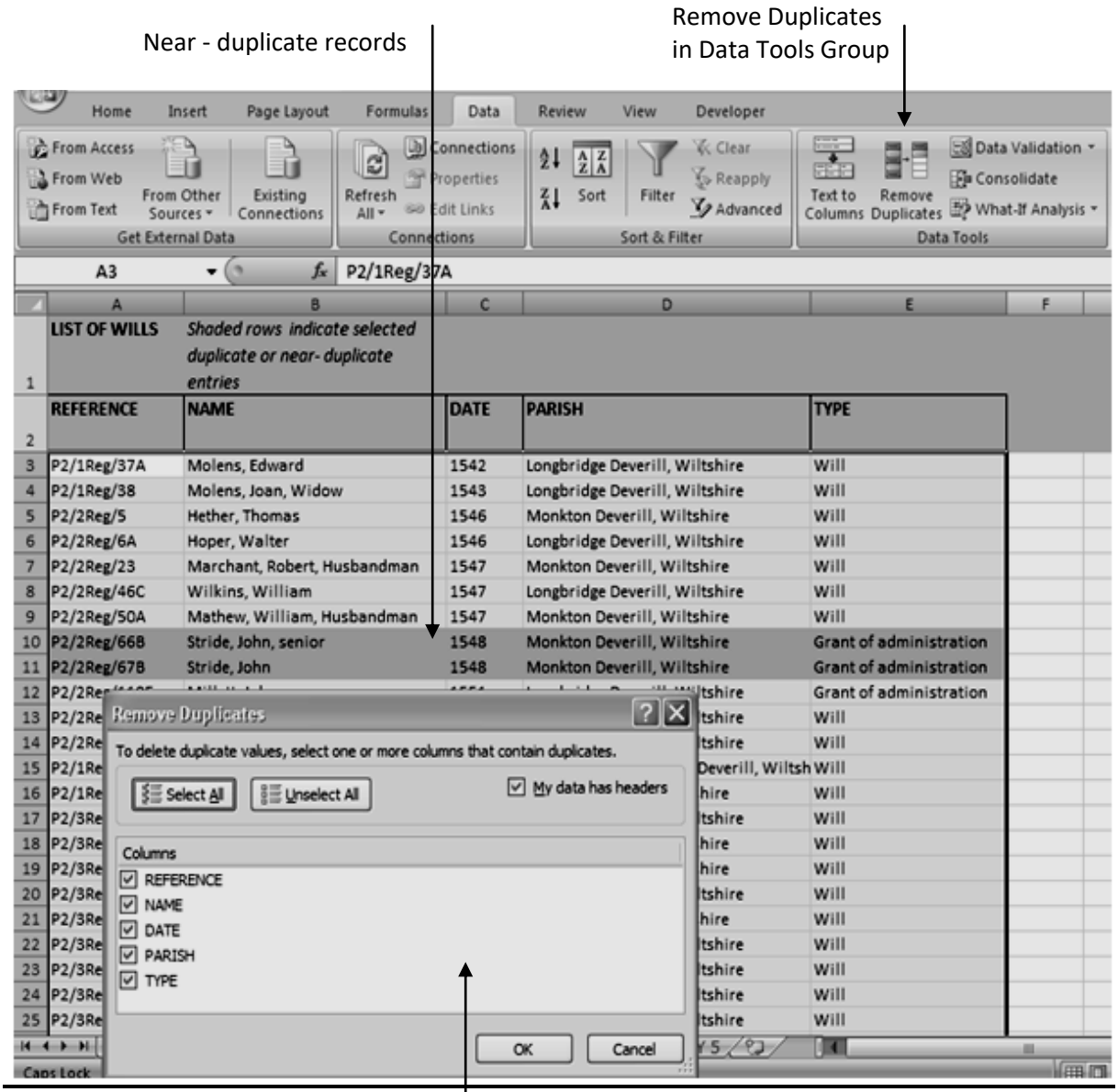
In the BLUEBERRY Workbook, use the worksheet entitled Copy 3, Copy 4 or Copy 5. Divide the Reference into just two separate columns. The first should contain the first three characters (P2/) and the second should contain the remainder (1Reg/37A).

- ***HINT:*** *In the Convert Text to Columns Wizard Step 1, choose the 'Fixed Width' button. The preview in Step 2 then offers the option to insert column breaks and drag them to the desired width.*

<< END OF THIS UNIT >>

UNIT G3
REMOVING DUPLICATE RECORDS

Purpose: To remove duplicate records using the 'remove duplicates' tool



A REMOVING EXACT DUPLICATES

➔ *Open the* **CHERRY** *Workbook*

1 Open the CHERRY Workbook
- *This contains a list of probate records*
- *The workbook contains several copies of the list (worksheets Copy 1 to Copy 5)*

2 View the worksheet entitled 'Original Data'.
- *The sheet contains the data before any transformations are carried out*
- *It also serves as a backup copy*
- *There are exactly 600 entries (rows 3 – 602)*

3 Observe that the list includes some duplicate and near-duplicate entries; a few have been shaded
- *For example, rows 10 and 11 are duplicates except for 'Reference' and 'Name'*
- *Rows 36 and 37 are duplicates apart from 'Type'*
- *Rows 42 and 43 are exact duplicates in all columns*

4 Open the worksheet entitled 'Copy 1'
Click in cell A1
Go to the Data Ribbon and locate the Data Tools Group
Click on Data Ribbon > Data Tools Group > Remove Duplicates
- *Remove Duplicates Dialogue Box opens*

5 In the Remove Duplicates Dialogue Box ensure that 'My Data Has Headers' is ticked
- *List of the column heading titles appears*
- *A tick appears next to each column title*

6 Click OK
- *Message appears as follows:*

"154 duplicate values found and removed; 446 unique values remain"

7 Click OK to close the message
Scroll down the list of probate records to examine it
If you wish, compare with the worksheet entitled 'Original Data' .
- *The list ends at row 448 (ie only 446 entries remain of the original 600)*
- *Rows in which the data is duplicated in every column (for examples, row 43) have been removed*

B REMOVING PARTIAL DUPLICATES

8 Open the worksheet entitled 'Copy 2'
Click in cell A1
Click on Data Ribbon > Data Tools Group > Remove Duplicates
- *Remove Duplicates Dialogue Box opens*

9 In the Remove Duplicates Dialogue Box ensure that 'My Data Has Headers' is ticked

10 In the Remove Duplicates Dialogue Box ensure that only the 'Type' box is ticked
Click OK
- *Message appears as follows:*

*"**577 duplicate values found and removed; 23 unique values remain"***

11 Click OK to close the message
Scroll down the list of probate records to examine it
If you wish, compare with the worksheet entitled 'Original Data' .
- *The list ends at row 24 (ie only 22 entries remain)*
- *All rows in which the 'type' was duplicated have been removed, leaving only a list of unique 'types'*
- *Some data entry errors in the original have been treated as 'unique values': see for example 'Will' and 'Wil' in rows 3 and 5; 'Account' and 'Accounts' in rows 14 and 16 and 'Commission' and ' commission' [with a leading space] in rows 17 and 18*

C REMOVING DUPLICATES FROM SELECTED ROWS AND MATCHING DATA IN MORE THAN ONE COLUMN

12 Open the worksheet entitled 'Copy 3'
Select rows 9 to 19

13 Click on Data Ribbon > Data Tools Group > Remove Duplicates
In the Remove Duplicates Dialogue Box ensure that 'My Data Has Headers' is NOT ticked

14 In the Remove Duplicates Dialogue Box tick boxes 'column C' and 'columns D' only (Columns C and D contain the date and the parish)

15 Click OK
- *Message appears as follows:*

2 duplicate values found and removed; 9 unique values remain

16 Click OK to close the message
Scroll down the list of probate records to examine it
Compare with the Original Data Worksheet if you wish

- *Duplicates have only been searched for in rows 9 to 19*
- *Duplicate row have been removed only if the data in columns C and D (date and parish) matched*

D REMOVING DUPLICATES : notes

17
- *Remove Duplicates is not case- sensitive*
- *Undo can reverse the effects of remove duplicates (but don't rely on it; take a back up copy first!)*
- *Remove Duplicates may not identify identical values formatted as different data types. For example, 1558 formatted as text and 1558 formatted as a number may be interpreted as different values*
- *Always do a 'reality check' after using Remove Duplicates*

18 Remove Duplicates can be used to remove multiple blank rows, since each empty row is a duplicate. The first empty row encountered will remain, but can easily be deleted manually

E IDENTIFYING DUPLICATES

19 'Remove duplicates' cannot be used to identify duplicates without removing them, but this can be done using a formula. Note that the following formula uses a conditional function (see UNIT J1) and absolute cell references (see UNIT H2), but you do not need to understand how it works in order to use it.

20 The following formula tells you how many copies there are of the data in a particular cell. It is essential to include the $ signs and that the two parts of the formula are identical. In this case, the formula looks for duplicates in cells C1 to C20. Enter the formula into cell F1, copying it exactly. Copy and paste it into cells F2 to F20.

=**COUNTIF($C$1:$C$20,$C$1:$C$20)**

REMOVING DUPLICATES

FURTHER PRACTICE

➔ *Open the **CHERRY** Workbook* which contains a list of probate records

Exercise 1

Use the worksheet entitled 'Copy 4'. Correct one or more of the data entry errors identified on Copy 2 (see section B, steps 8 – 11 above). You might identify and correct the errors by filtering the original data or by using find and replace. Then remove duplicates and compare the results with those obtained on Worksheet 2.

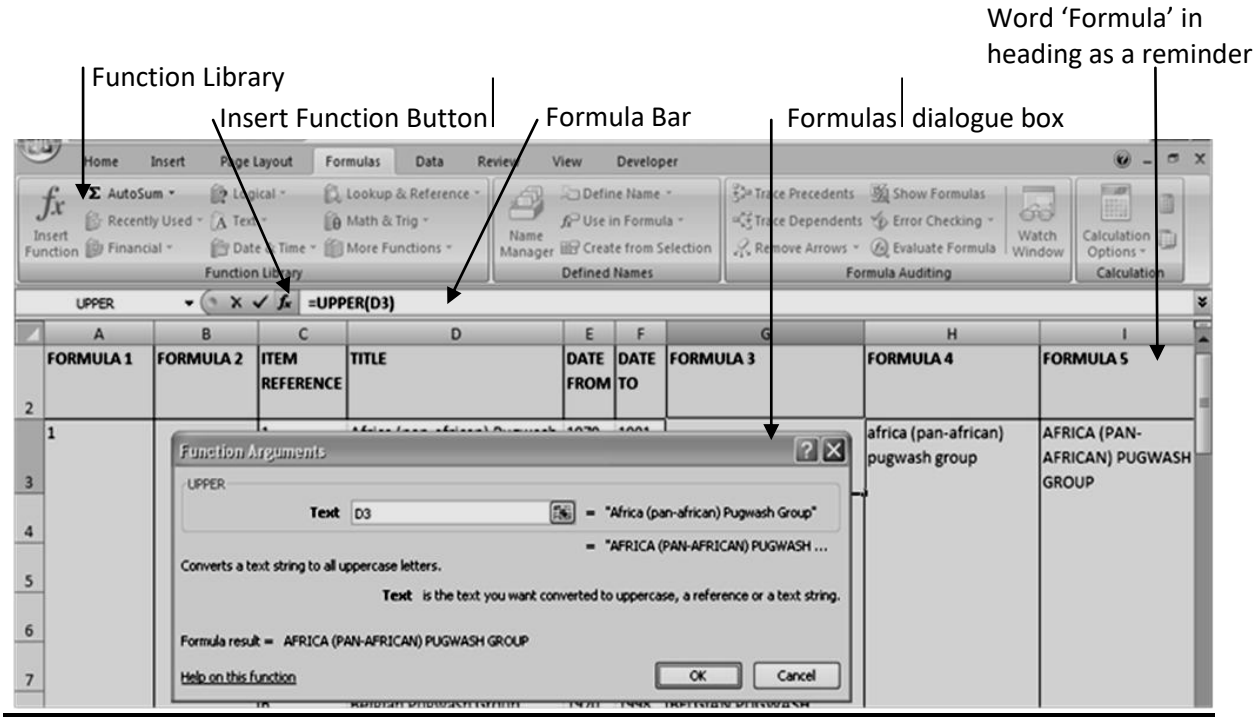
Exercise 2

Use worksheets entitled 'Copy 4' and 'Copy 5' to experiment with removing duplicates, for example, compare the results when ticking different combinations of columns in the 'Remove Duplicates' Dialogue Box.

<< END OF THIS UNIT >>

UNIT H1
TECHNIQUES FOR WORKING WITH FORMULAS

Purpose: To summarise some techniques to assist with creating, copying, viewing and checking formulas.



A INTRODUCTION

➔ *Open the GOOSEBERRY Workbook*

1 This unit covers the following topics relating to entering and using Functions and Formulas:
Creating formulas (Section B)
Copying formulas (Section C)
Viewing and checking formulas (Section D)

2 This unit uses the GOOSEBERRY Workbook. This contains an extract from a catalogue, columns intended for insertion of formulas, and some completed formulas.

- *You may OMIT the practical exercises if you are already familiar with the techniques described*

3 For an introduction to using Functions and Formulas, see the following units:
UNIT E1 Introducing Functions and Formulas
UNIT E2 Using Functions and Formulas

B CREATING FORMULAS

4 Formulas can be created in a variety of ways including:
Method 1: Using the Formulas Ribbon > Function Library to open a dialogue box for a particular function
Method 2: Using the 'Insert function' button next to the Formula Bar to open a dialogue box for a particular function
Method 3: If you know the function and the syntax it requires, the formula can be typed direct into the cell or the Formula Bar
Method 4: Copy and paste a previous-created formula (see section C below)

5 *Practical exercise 1*
Open the Gooseberry Workbook
Go the worksheet entitled Exercise 1

- *You may OMIT the practical exercises if you are already familiar with the techniques described*

6 ***Method 1: Using the Function Library***
Select cell G3 (column 'Formula 3') by clicking in it
Use the Function Library (Formulas Ribbon) to select the function 'UPPER'
- *Dialogue Box appears*

7 In the dialogue box, enter D3 (or click in cell D3 to select it)
Click 'OK'
- *The formula* **=UPPER(D3)** *is inserted into cell G3 (see the formula bar)*
- *The contents of cell D3 are displayed in cell G3, in Upper Case*

8 ***Method 2: Using the Insert Function Button***
Ensure that cell G3 contains upper case text (as created in steps 6 and 7 above)
Ensure that cell H3 is blank (containing no formula nor data)
Select cell H3 (column 'Formula 4') by clicking in it
Click the Insert Function button (**Fx** to the left of the Formula Bar)
- *Dialogue Box appears*

9 In the 'Search for a function' box enter 'lower case'
Then click 'GO'
- *List of functions appears in response to the search*

10 From the list of functions select 'LOWER'
Click 'OK'
- *Dialogue Box appears*
- *Note that this is exactly the same dialogue box as appears if the function is selected from the Function Library*

11 In the dialogue box, enter G3 (or click in cell G3 to select it)
Click 'OK'
- *The formula* ***=LOWER(G3)*** *is inserted into cell H3 (see the formula bar)*
- *The contents of cell D3 are displayed in cell H3, in Lower Case*
- *Note that the function is acting on the* ***formula*** *in cell G3, not on the original text in cell D3*

12 ***Method 3: Type the formula***
Select cell I3 (column 'Formula 5') by clicking in it
Type **=UPPER(H3)**
- *The contents of cell D3 are displayed in cell I3, in Upper Case*
- *Note that the function is acting on the* ***formula*** *in cell H3, not on the original text in cell D3*

C COPYING FORMULAS

13 Formulas can be copied and pasted in a variety of ways including:

- *Method 1: Copy and Paste*
- *Method 2: Dragging the Autofill handle*
- *Method 3: Clicking the Autofill Handle*

14 *Practical exercise 2*
Open the GOOSEBERRY Workbook
Go the worksheet entitled Exercise 2

- *This contains the formulas created in Exercise 1*
- *You may OMIT the practical exercises if you are already familiar with the techniques described*

15 ***Method 1: Copy and Paste***
Select cell G3 (column 'Formula 3') and copy it
Select cell G4 and a few cells below (down to about G16)
Paste the cell contents

- *The formula (=UPPER...) appears in the selected cells*

16 ***Hint***
To select multiple cells before pasting, use the NAME box:
Click in the NAME box (to the left of the formula bar)
Type the range of cells to be selected (eg G4:G16 selects all of cells C4 to C16)

- *The chosen cells are greyed to show they have been selected*

17 ***Method 2: Dragging the Autofill handle***
Select cell H3 (column 'Formula 4')
Use the autofill handle to drag the formula to some of the cells below in column H (down to about H12)

- *The formula (=LOWER...) appears in the selected cells*

18 ***Method 3: Clicking the Autofill handle***
Select cell I3 (column 'Formula 5')
Double click on the autofill handle

- *The formula (=UPPER...) appears in some cells in columns I*
- *It ONLY fills down to however far you have inserted content into column H. Once Excel finds a blank cell in column H, it stops.*

D VIEWING AND CHECKING FORMULAS

19 This section covers the following techniques for viewing and checking formulas:
Method 1: Viewing a formula and its dependent cells by double clicking in the cell
Method 2: Viewing a formula and its dependent cells by clicking in the Formula Bar
Method 3: Using a keystroke to view all formulas within cells
Method 4: Re-open the dialogue box

20 **Viewing and Checking Formulas: Practical Exercise**
Open the GOOSEBERRY Workbook
Go the worksheet entitled Exercise 3
- *This includes the formulas created in Exercise 2*
- *It also includes some additional formulas in column J ('Formula 6'). Note that you do not need to understand the detail of these additional formulas*

21 ***Method 1: Double click within Cell***
Click in cell J5 (or any other cell containing a formula)
- *Notice that the formula appears in the FORMULA BAR, but in the cell you have clicked in the DATA appears*

22 DOUBLE-click in cell J5 (or any other cell containing a formula)
- *The formula displays within the cell as well as within the formula bar*
- *The cells which it acts on are indicated with coloured borders*
- *In the cell (but not in the formula bar) the cell references within the formula are coloured to match the borders of the appropriate cells*

23 ***Method 2: Click in Formula Bar***
Click in cell J6 (or any other cell containing a formula)
Click in the formula bar
- *The formula displays within the cell as well as within the formula bar*
- *The cells which it acts on are indicated with coloured borders*
- *In the formula bar (but not in the cell) the cell references within the formula are coloured to match the borders of the appropriate cells*

24 ***Method 3: Use a keystroke to view all formulas within cells***
Click in any cell on the worksheet
Hold down the ctrl key
Press the key to the left of the 1 on the keyboard, containing the ` symbol (and probably also the ¬ and ¦)
- *Formulas appear in all the cells as well as in the formula bar*
- *To reverse the display, click control plus the key again.*

Method 4: Re-open the dialogue box

Click on cell G5

Click the Insert Function button (**Fx** to the left of the Formula Bar)

- *Dialogue Box appears*
- *The data in the dialogue box (in this case the cell reference D5) can be amended if required*

If a dialogue box is not available (for example, in the case of a complex formula), the result of the formula displays instead:

Click in cell J5

Click the Insert Function button

- *Function Arguments Box appears instead of a dialogue box*

E FUNCTIONS: THINGS TO REMEMBER

(These notes are repeated from UNIT E2 for convenience)

The formula is only a rule: what you see in the 'formula' cell is NOT FIXED: it will change if the values to which the formula refers alter or are removed. If you want to fix the data use Copy > Paste Values

The formula is not normally activated until you click 'Enter'

'Formula' means the whole expression starting with the equals sign, such as ***=SUM(B3:B6)*** *; 'Function' is the special word which makes the formula work (in this case 'SUM').*

Shading formula cells or columns is a useful technique to help remind you that are working with a formula cell, not a fixed value. You can also use the word 'Formula' in the column heading.

You do not need to use the formulas dialogue box. If you are familiar with the function you want to use, simply type the formula into to the appropriate cell or in the formula bar.

Formulas will not work if you type them into a cell formatted as 'text'. Change the cell format to 'number' format before starting to enter a formula.

There are over 400 functions available in Excel, and most can be combined to produce extremely complex expressions. If you can formulate it logically, Excel can probably do it!

Use Help to find the function you need

The best way to find out what a particular function does is to experiment

<< END OF THIS UNIT >>

UNIT H2
RELATIVE AND ABSOLUTE CELL REFERENCES

Purpose: To explain how Excel interprets cell references

A SUMMARY

1 Cell references in the form **A3** [column A row 3] are RELATIVE: if you include a cell reference in a formula, the cell references will automatically change when you move or copy the formula to a new cell.

2 Cell references in the form **$A$3** are ABSOLUTE: they remain unchanged when you move or copy a formula to a new cell.

3 MIXED cell references are possible: **$A3** indicates that the column (A) is absolute but the row (3) is relative; in the format **A$3** the column is relative and the row is absolute.

4 Both relative and absolute references adjust automatically when new rows or columns are inserted.

B PRACTICAL EXAMPLES

➔ *Open the GOOSEBERRY Workbook*

5 These exercises use the GOOSEBERRY Workbook
Go the worksheet entitled Exercise 4
- *This contains an extract from a catalogue and columns containing formulas (for step by step instructions on these formulas see UNIT H1).*
- *You may OMIT the practical exercises if you are already familiar with the concepts described*

6 **Relative cell references**
Click in cell H3 and look at the formula in the formula bar
- *The formula is* **=LOWER(G3)**

7 Copy the formula in cell H3 and paste into cells H4 to H7
Click in cell H7 and look in the formula bar at the formula you pasted in
- *The formula is* **=LOWER(G7):** *the cell reference has changed relative to the cell in which the formula appears*

8 **Using relative cell references**
In Cell E1 type any number (eg 24)
Press ENTER

9 Click in cell A3 (column 'Formula 1)
View the formulas which it contains
- *Cell A3 contains the formula: =C3+E1*
- *This generates a new item reference, higher than the original by whatever value you entered in E1*

10 Copy the formula in cell A3 to a few cells in column A (down to about A7), and view the results
- *Instead of adding your chosen number to the item reference, the formulas add the item reference and the 'date from' (including in cell E4, trying to add the item reference to the words 'date from', resulting in an error message)*
- *This is because E1 has changed relative to the cell in which the formula appears in column A*

11 **Using absolute cell references**
Click in cell B3 (column 'Formula 2)
In cell B3 use any method to enter the following formula: **=C3+$E$1**
- *This is identical to the formula in cell A3, except for the $ signs*
- *This generates a new item reference, identical to that in cell A3*

12 Copy this formula down a few cells in column B, and view the results
- *All the formulas add the value in cell E1 to the item reference; adding the dollar signs creates an **Absolute Reference***
- *The reference number used (in column C) changes as appropriate, as the column C reference is a **Relative Reference***

13 Change the number in cell E1. Then click Enter.
Notice the values in column B
The values in column B adjust accordingly

14 **Behaviour of formulas when inserting new rows or columns**
Insert a new row and/or column, so that the number you entered in cell E1 is now in cell F1 or E2 or F2
Note what happens to the formulas in columns A and B
- *Both absolute and relative references adjust if new rows or columns are inserted*
- *The results of the formulas in column B do not change (but the formulas themselves in column B have all adjusted to refer to the correct cells)*

RELATIVE AND ABSOLUTE CELL REFERENCES
FURTHER PRACTICE

Using the GOOSEBERRY Workbook, worksheet entitled 'more practice', experiment with the effects of using relative and absolute cell references in formulas. Or create your own data if you wish.

Depending on what you want to focus on, you could include some of the following:

Different types of cell reference

- **Relative** cell references (in the form **A1**)
- **Absolute** cell references (**$A$1**)
- **Mixed** cell references (**$A1** or **A$1**)

Different types of formula

- A simple formula such as **=A1** may be quickest to enter
- Calculations (such as adding a number of years to a date (eg **=C3+25**) may help to identify circumstances generating error messages
- Text-based formulas such as **=LEFT(A1)** may be less likely to result in error messages

Copying formulas to different places

- Copy a formula and paste to successive ROWS or successive COLUMNS: look at the formula in the formula bar, to observe what changes (if any) Excel makes to your original formula.

Inserting rows and columns

- Copy a formula into successive rows or columns, then insert new rows or columns within the data. Before and after inserting columns and rows, look at the formulas in the formula bar, to observe what changes Excel makes to them.

<< END OF THIS UNIT >>

UNIT H3
USEFUL FUNCTIONS AND SYMBOLS

Purpose: To summarise useful functions and formulas
This unit is for reference and contains no practical exercises

The following functions and symbols are likely to be of particular use when creating formulas for data cleaning.

A FUNCTIONS WHICH CHANGE TEXT

Function	Description	Example
LEFT	Extracts a given number of characters, counting from the left	=LEFT(A1,3)
RIGHT	Extracts a given number of characters, counting from the right	=RIGHT(B1,6)
MID	Extracts a given number of characters, starting at a chosen character	=MID(C1,5,7)
LOWER	Converts text to all lower case	=LOWER(B1)
UPPER	Converts text to all upper case	=UPPER(B1)
PROPER	Converts the first character in each word to upper case and the remainder to lower case	=PROPER(B1)
CONCATENATE	Joins the contents of two or more cells and/or typed text.	=CONCATENATE(A1,"and",B1)
VLOOKUP	Changes one value to another using a look up list (*see UNIT J3 Transforming data using a look-up list*)	=VLOOKUP(H4,'list!$A$1:$B$12,2,FALSE)
CLEAN	Removes non-printing characters	=CLEAN(D1)
TRIM	Removes excess spaces, leaving only a single space between words	=TRIM(D1)

B FUNCTIONS WHICH PROVIDE INFORMATION

Function	Description	Example
COUNTA	Counts the number of non-blank cells in a given range	=COUNTA(A1:M10)
COUNTIF	Counts the number of cells in a given range which meet a specified criterion	=COUNTIF(D2:D99,"ant")
COUNTIFS	Counts the number of cells meeting multiple criteria in multiple ranges	=COUNTIFS(D1:D99,"the",E1:E99,"or")
FIND	Finds the position of a particular text string (word, phrase or letter) within a cell. Case sensitive. Error message is returned if a match is not found	=SEARCH("inv",F3,1)
SEARCH	Finds the position of a particular word or letter within a cell (not case sensitive). Error message is returned if the word is not found	=FIND("inv",F3,1)
LEN	Finds the length of a text string	=LEN(F3)

C CONDITIONAL AND LOGICAL FUNCTIONS

Function	Description	Example
IF	See UNIT J1 Conditional Formulas	=IF(B5>7,"more than 7”)
IFERROR	Used to prevent error messages displaying. (*See UNIT J2 Combining Formulas*)	=IFERROR(A1,"")
AND	Compares two or more statements. If all are true, returns TRUE, otherwise returns FALSE	=AND(B3>6,B4<10)
OR	Compares two or more statements. If at least one is true, returns TRUE, otherwise returns FALSE	=OR(B3>6,B4<10)
TRUE	Enables a formula to use the results of a function which returns TRUE or FALSE	=IF(A1=TRUE,"yes","no")
FALSE	Enables a formula to use the results of a function which returns TRUE or FALSE	=IF(A1=FALSE,"no","yes")

D MATHEMATICAL AND OTHER SYMBOLS

The following symbols can be used in formulas when working with text as well as with numbers

Symbol	Description	Example
=	At the start of a cell indicates a formula. In any other position, acts as 'equals'. Returns TRUE if the terms match (not case–sensitive)	ANT=ant
<>	Does not equal. Returns TRUE if the terms do not match (not case–sensitive)	ANT<>BEE
>	Greater than. For text, is interpreted as 'later alphabetically'	Bee>ant
<	Less than. For text, is interpreted as 'earlier alphabetically'	Ant<bee
>=	Greater than or equal to (the equals sign must come second)	Bee>=ant
<=	Less than or equal to (the equals sign must come second)	Ant<=bee
&	Joins the contents of two or more cells and/or typed text (similar effect to the function CONCATENATE)	=A1&"and"&B1
:	Used with cell references to define a range of cells	A1:M67
+ -	Plus and minus signs can normally only be applied to numbers. Using them with text normally results in an error message	

<< END OF THIS UNIT >>

UNIT J1
CONDITIONAL FORMULAS

Purpose: To transform data in different ways by applying a rule

A INTRODUCTION

➔ *Open the LEMON Workbook*

1 Most functions return a result according to a single rule.
For example **=LEFT(C4,3)** always returns the three leftmost characters in cell C4

2 Conditional statements return a choice of results, depending on whether or not specified conditions are met.
For example, one rule is followed if the cell contains a number greater than 3, and a different rule if it contains instead the word 'banana'.

3 Conditional formulas operate by evaluating whether a particular condition is met or not. Excel expresses this using the functions TRUE and FALSE.

4 The practical examples use the LEMON Workbook

- *NOTE: in the practical examples, the Formulas Ribbon> Function Library is always used to open a dialogue box. If you prefer, use the 'insert function' button next to the Formula Bar instead.*

B SIMPLE EXAMPLE

5 Open the LEMON Workbook, Worksheet 'Introduction'
In cell E7 type your name
In cell E8 Type your name again

6 Click in Cell E9
From the function library on the Formulas Ribbon choose 'Logical'
From the drop down list, select IF

- *A dialogue box appears*

7 In the first box ('Logical test') enter **E7 = E8**

- *The word 'TRUE' appears to the right of the 'Logical Test' box (because the contents of cells E7 and E8 are - or should be – identical)*

8 In the second box (value if true) type the words **The Same**
In the third box (value if false) type the words **Different**

- *Note that the 'value if true' and the 'value if false' can be any word, phrase or number – or even a formula*

9 Click OK

- *Dialogue Box closes*
- *The value **The Same** appears in cell E9*

10 Change the value in cell E7 to your surname
Click enter (or click in a cell other than E7)

- *The value in cell E9 changes to **Different***

11 Change the value in cell E8 to your surname
Click enter (or click in a cell other than E7)

- *The value in cell E9 changes back to **The Same***

12 Click in Cell E9
Look at the Formula Bar and examine the formula which the dialogue box has generated

- *The formula reads **=IF(E7=E8,"The Same","Different")***
- *Notice that the two phrases entered as the 'value if true' and the 'value if false' appear in double quotes. Excel has recognised them as text, and uses the double quotes to prevent trying to interpret them as a function (which might happen for example, if you had entered the word 'left').*

13 REMEMBER that just like any other formula, those using the IF function are only rules. What you see in the 'formula' cell is not fixed and will change if the values to which the formula refers alter or are removed. If you want to fix the data use Copy > Paste Values

C USING 'IF' WITH A SPECIFIC VALUE

14 Open the LEMON Workbook, Worksheet 'Wills copy 1'

- *This contains a list of probate records*
- *In this exercise, IF will be used to identify records dated 1578*

15 Click in Cell E4 (column 'date formula 1')
From the function library on the Formulas Ribbon choose 'Logical'
From the drop down list, select IF
- *A dialogue box appears*

16 In the first box ('Logical test') type **D4=1578**
- *The word 'FALSE' appears to the right of the 'Logical Test' box (because the date in D4 does not equal 1578)*

17 In the second box (value if true) type **Yes**
In the third box (value if false) type **No**
- *Remember that the 'value if true' and the 'value if false' can be any word, phrase or number*

18 Click OK
- *Dialogue Box closes*
- *The value 'No' appears in cell E4*

19 Copy the formula to the rest of column E
- *Cells where the date is 1578 display 'yes' in column E*
- *Most cells in column E display the value 'No' (because the date does not equal 1578)*

D USING MATHEMATICAL OPERATORS TO COMPARE VALUES

20 Open (or continue using) the LEMON Workbook Worksheet 'Wills copy 1'
- *In this exercise, IF will be used to identify records dated later than 1600*

21 Click in Cell F4 (column 'date formula 2')
From the function library on the Formulas Ribbon choose 'Logical'
From the drop down list, select IF

22 In the first box of the dialogue box ('Logical test') enter **D4>1600**
- *The word 'TRUE' appears to the right of the 'Logical Test' box (because the date in D4 is indeed greater than 1600)*

23 In the second box (value if true) type **Greater**
In the third box (value if false) type **No**
- *Remember that the 'value if true' and the 'value if false' can be any word, phrase or number*

24 Click OK
- *Dialogue Box closes*
- *The value 'Greater' appears in cell F4*

25 Copy the formula to the rest of column F
- *Cells in column F display either the value 'no' (where the date is not greater than 1600) or 'Greater' (where the date is greater than 1600)*

26 Mathematical symbols which can be used include <= (less than or equal); <> (not equal).
- *For a list of mathematical symbols, see section D of UNIT H3 Useful functions and symbols.*

E USING 'IF' TO COMPARE VALUES IN TWO CELLS

27 Open (or continue using) the LEMON Workbook Worksheet 'Wills copy 1'
- *In this exercise, IF will be used to identify records later in date than the value entered in another cell*

28 Click in Cell G4 (column 'date formula 3')
From the function library on the Formulas Ribbon choose 'Logical'
From the drop down list, select IF

29 In the first box of the dialogue box ('Logical test') enter **D4<$D$2**
- *Note that the $ signs in $D$1 make this an **absolute** cell reference, which will not change when copied [see UNIT H2 Understanding relative and absolute cell references]*
- *The word 'FALSE' appears to the right of the 'Logical Test' box (because the date in D4 is not less than the value in cell D2)*

30 In the second box (value if true) type **Earlier**
In the second box (value if false) type "" [with no space between the pair of double quote marks; this means nothing is to display if the result of the formula is false]
Click OK to close the dialogue box

31 Copy the formula to the rest of column G
- *Cells in column G display the value 'earlier' where the date is earlier than that in cell D2*
- *Cells in column G display nothing where the date is not earlier than that in cell D2*

32 Change the value in cell D2
- *Notice that the values displayed in column G change accordingly*
- *Although not included in this unit, note that it is now possible to apply a filer [See UNIT D2] to display only the 'earlier' records*

F USING 'IF' IN COMBINATION WITH ANOTHER FUNCTION

33 The 'logical test' for the IF function can be complex, and can incorporate other functions. In the exercise which follows, the last six characters of the name (column B) are extracted using the function RIGHT. If they are equal to the word 'senior' then XXXX is displayed; otherwise nothing displays
- *For more about combining functions and formulas, see UNIT J2 Combining Formulas*

34 Open (or continue using) the LEMON Workbook, Worksheet 'Wills Copy 2'
Click in Cell E4 (column 'Senior formula')

35 Open the 'IF' dialogue box
In the first box ('Logical test') type **RIGHT(B4,6)="Senior"**
- *The first part **RIGHT(B4,6**) is almost identical to the formula =RIGHT(B4,6) which would return the last six characters of cell B4, but omitting the initial equals sign*
- *The second part **="Senior"** asks Excel to test whether the result of the formula =RIGHT(B4,6) matches the word 'senior'*
- *The word Senior MUST be enclosed in double quotes to indicate that it is text*

36 In the second box (value if true) type XXXX
In the second box (value if false) type "" [with no space between the pair of double quote marks]
Click OK to close the dialogue box

37 Copy the formula to the rest of column E
- *Cells in column E display the value XXXX where the last word of the name is 'senior', others nothing is displayed*

38 Apply a filter to column E
Use it to display only those rows with the value XXXX
- *This formula has made it possible to filter on only those names ending in 'senior' (for example, to edit them before dividing the data into columns for first and second name)*

G OTHER CONDITIONAL FUNCTIONS

39 There are a number of other conditional functions: these can be seen from the 'Logical' menu on the function library. The most useful are probably AND, OR and IFERROR.

40 AND combines two or more separate conditions which must all be true for the rule to take effect
OR combines two or more separate conditions only one of which needs to be true for the rule to take effect

- *AND and OR return either the values TRUE or FALSE, but can be combined with IF to produce a more complex result.*

41 The conditional function IFERROR can be used to prevent display of error messages. See UNIT J2 *Combining Formulas* for examples of its use

42 For more functions, see UNIT H3 *Useful functions and symbols.*

43 The best way to understand the effect of different functions is to experiment

- *Don't forget to take a back up copy first!.*

CONDITIONAL FORMULAS
FURTHER PRACTICE

Exercise 1

If you would like additional practice on the activities in this unit, repeat them using the LEMON Workbook, Worksheet 'Wills Copy 3'; you could experiment with introducing variations such as changing the logical test or the contents of the 'value if true' and 'value if false' boxes.

Exercise 2

Use the LEMON Workbook, Worksheet 'More Practice'. By entering a conditional formula in column E, filter the data to display only those rows where the start date is later than the end date (indicating a data entry error)

Exercise 3

Use the LEMON Workbook, Worksheet 'More Practice'. By entering a conditional formula in column E, filter the data to display only those rows where the first word of the Title is 'Gloucester'.

Exercise 4

Use the LEMON Workbook, Worksheet 'More Practice'. By entering a conditional formula in column E, filter the data to display only those rows where the first letter of the Title is later alphabetically than F

- *Hint: the letter F must be enclosed in double quotes (* **>"F"** *).*

Exercise 5

If you are already confident using conditional statements, experiment with using AND and OR.

<< END OF THIS UNIT >>

UNIT J2
COMBINING FORMULAS

Purpose: To use different functions and formulas together

A INTRODUCTION

1 There are two ways of combining formulas:
Method 1: multiple formulas in separate columns
Method 2: a single complex formula

2 **Summary of method 1: multiple formulas**

- Enter each formula in a separate column. The first acts on the original data, and subsequent formulas act on the result of earlier formulas until the end result is achieved
- This method is relatively easy to use, as the appropriate dialogue boxes can be used to enter each formula.
- Each of the formulas can be altered independently without affecting the others
- For data cleaning purposes, where the project is intended to create a single set of new data, this is likely to be the most appropriate method

3 **Summary of method 2: a single complex formula**

- Only one column is used, containing a single complex expression which combines all the steps required
- This method is time-consuming and prone to error: the syntax, including nested brackets, must be exactly correct
- Dialogue boxes cannot normally be used
- Changing one part of the expression is likely to introduce errors to other parts, which can be hard to spot
- This method is most appropriate when creating sophisticated Excel documents, such as forms and templates for frequent use, where multiple columns of formulas may be intrusive
- This method is not normally necessary for data cleaning. However, it is covered in optional section C below for those who wish to attempt it.

4 **Combining methods 1 and 2**
With experience, it may sometimes save time to combine some very simple formulas into a single more complex expression, even when using method 1

B PRACTICAL EXAMPLE (Method 1: multiple formulas)

➔ *Open the PLUM Workbook*

5 Open the PLUM workbook, Worksheet 'Copy 1'

- *This contains a list of probate records.*

6 The aim of this exercise is to identify which records contain an inventory
Examine the data in column F (TYPE)

- *This contains a number of records for 'inventory'*
- *The word 'inventory' appears in different contexts (for example, as a single word, or at the start, end or in the middle of a list)*

7 Examine the formula which has been entered in column G (FORMULA 1):
Click in cell G3
View the formula in the Formula bar

8 Click the Insert Function button (**Fx** to the left of the Formula Bar) in order to view the dialogue box which was used to create the formula in cell G3

- *This formula uses the function 'SEARCH' which returns the position of the word 'inv' in the 'type' column*
- *For example, in cell F3, the word starts at the seventh character, so 7 is displayed in cell G3*

9 Click OK to close the dialogue box

- *Where the word 'inv' is not found, excel returns an error message (#VALUE!)*
- *The abbreviation 'inv' has been chosen rather than the full word in case the column contains mis-spellings or plurals*

10 Examine the formula which has been entered in column H (FORMULA 2)
Click in cell H3
View the formula in the Formula bar
Click the Insert Function button (**Fx** to the left of the Formula Bar) in order to view the dialogue box which was used to create the formula in cell H3

- *This formula uses the conditional function IF*
- *It acts on the value created by the formula in column G: if the value is greater than 0, then the word 'inventory' is displayed*

11 Click OK to close the dialogue box

- *Where column G contains an error, this is carried forward to column H*
- *Despite the 'error' it is still possible to filter in order to display only the rows including an inventory*
- *The error would be inconvenient if the data were to be copied and pasted to produce a clean copy*

12 Examine the formula which has been entered in column I (FORMULA 3)
- *This formula uses the CONCATENATE function to combine the word 'inventory' with the date (column D)*
- *Again, the error value persists: this may not matter if the intention is to delete any records not containing inventories.*

13 The following steps use the **IFERROR** conditional function to remove the error values.

14 Click in cell J3 (column FORMULA 4)
From the function library on the Formulas Ribbon choose 'Logical'
From the drop down list, select IFERROR
- *A dialogue box appears*

15 In the first box ('Value') type **I3**
- *If there is no error, then the value in J3 will be the same as in I3*

16 In the second box (Value if Error) type **""**
- *If there is an error, nothing is to display*

17 Click OK to close the dialogue box
Copy the formula to the rest of column J
- *Cells with no error display the same value in column I and column J*
- *Cell with an error in column I display no data in column J*

C PRACTICAL EXAMPLE (Method 2: complex formulas)

➔ *Open the **PLUM** Workbook*

18 **This section is OPTIONAL.**
It is technically quite demanding, and it is recommended that you study this section **only** if you are experienced with using functions and formulas.

If you do not wish to study combining formulas into a single complex formula, or do not have time at present, omit it and go straight to the practice exercises at the end of this Unit.

19 If you simply wish to see an example of a complex formula, open the PLUM workbook, Worksheet 'Copy 2'.
Examine the formula which has been entered in column M (Combined 3)
Click in cell M3
View the formula in the Formula bar

- *The complete formula is =IFERROR(IF(SEARCH("inventory",F3,1)>0,CONCATENATE(H3,", ",D3),""),"")*
- *This has exactly the same effect as the four separate formulas created in section A (columns Formula 1,2,3,4)*

20 The following steps build up this formula in stages
Open the PLUM workbook, Worksheet 'Copy 2'
This contains a list of probate records, and the formulas created in section B above.

21 Click in cell K3 (column COMBINED 1)
From the function library on the Formulas Ribbon choose 'Logical'
From the drop down list, select IF

22 In the dialogue boxes enter data as follows:
Logical Test **SEARCH("inv",F3,1)>0**
Value if true **"inventory"**
Value if false **""**
Click Ok to close the dialogue box

- *the value 'inventory' displays in cell K3*

23 Copy the formula to the rest of column K

- *Where column F contains the word 'inv' then the word 'inventory' is displayed*
- *Where the word 'inv' is not found, Excel returns an error message (#VALUE!)*

24 Click in cell K3
View the formula in the Formula bar
The complete formula reads **=IF(SEARCH("inv",F3,1)>0,"inventory","")**

- *Note that the brackets are essential to ensure that Excel interprets the formula correctly*

25 Compare the formula in column L with those in columns G and H
Either view the formulas in the Formula bar
Or click the formula button and view the dialogue boxes

- *'Value if true' and 'value if false' are the same in column L as in column H*
- *In column G, the complete formula is* **=SEARCH("inventory",F3,1)**
- *In column H, the logical test is* **G3>0**
- *In column K, the formula from column G has replaced the cell reference G3*
- *The equals sign (=SEARCH...) is not repeated. This is only needed at the start of the whole combined formula*

Summary of the relationship between the formulas in columns G, H and L

	column G	**column H**	**Column L**
Function	SEARCH	IF	IF
Logical Test		G3>0	SEARCH("inv",F3,1)>0
Value if true		"inventory"	"inventory"
Value if false		""	""
Find Text	inv		
Within text	F3		
Start Number	1		
Complete formula	=SEARCH("inv",F3,1)	=IF(G3>0,"inventory","")	=IF(SEARCH("inv",F3,1)>0,"inventory","")

26 Examine the formula which has been entered in column L (Combined 2)
Click in cell L3
View the formula in the Formula bar

- *The complete formula is*
  =IF(SEARCH("inventory",F3,1)>0,CONCATENATE(H3,", ",D3),"")

27 Click the formula button and view the dialogue box which was used to create the formula in cell L3

- *This combines into a single expression the formulas in columns G (SEARCH), H (IF) and I (CONCATENATE)*
- *The Logical test is derived from column G:* **SEARCH("inv",F3,1)>0**
- *The value if true is derived from column I:* **CONCATENATE(H3,", ",D3)**
- *As in columns G, H and I, there is an error value if the word 'inv' is not found*

28 Examine the formula which has been entered in column M (Combined 3)
Click in cell M3
View the formula in the Formula bar

- *The complete formula is*
  =IFERROR(IF(SEARCH("inventory",F3,1)>0,CONCATENATE(H3,", ",D3),""),"")

29 Click the formula button and view the dialogue box which was used to create the formula in cell M3

- *This combines into a single expression the formulas in columns G (SEARCH), H (IF), I (CONCATENATE) and J (IFERROR)*
- *The Value is derived from column M:*
  IF(SEARCH("inventory",F3,1)>0,CONCATENATE(H3,", ",D3),"")
- *The value if error is the same as for column J: ""*

30 Hints on generating complex formulas

- *Test the effects of each separate element, and identify errors, by creating them first as independent formulas in separate columns.*
- *Combine formulas incrementally, rather than trying to create a single complex formula from scratch*
- *To reduce the possibly of making a syntax error, copy a formula which you know works and paste it into the dialogue box used to create the more complex formula (but don't forget to delete the initial = sign, which is only required at the start of the whole formula)*
- *It can be helpful to develop a complex formula outside Excel, for example by typing it in Word or a text editor, and pasting the completed formula into Excel.*
- *Formatting issues may mean Excel fails to recognise the pasted-in expression as a formula. If this happens, paste the formula into Excel without the initial equals sign. Then type the = to turn it into a formula.*

CONDITIONAL FORMULAS
FURTHER PRACTICE

Exercise 1

Open the PLUM Workbook, Worksheet 'Practice exercise 1'. Using similar steps to those in section B (finding those records with the word 'inventory' in column F) find those records with the word 'bond' in column F.

Exercise 2

Open the PLUM Workbook, Worksheet 'Practice exercise 1'. Using a sequence of formulas, find those records meeting the following conditions:

- *the word 'bond' in column F*
- *AND the word 'Reg' [indicating registered wills] in column A.*

Exercise 3

Open the PLUM Workbook, Worksheet 'Practice Exercise 3' which contains an extract from a catalogue. Use a sequence of formulas to identify those records where the number of characters in the title and description fields combined (columns C and D) is greater than 300 (this could be an essential first step if migrating to a database with limited space)

- *Hint: the function LEN returns the number of characters*

Exercise 4
(complex formulas)

Repeat any of exercises 1 to 3, but combining as many formulas as you can into a single complex formula.

<< END OF THIS UNIT >>

UNIT J3
TRANSFORMING DATA USING A LOOK-UP LIST

Purpose: To change one set of values to another using a list of equivalents (and the function VLOOKUP).

A INTRODUCTION

➔ *Open the* RASPBERRY *Workbook*

1 The function VLOOKUP transforms one set of values to another using a list of equivalents. For example, changing abbreviated month names Jan, Feb, Mar to complete month names January, February, March. The effect is similar to Find and Replace, but without having to change each term individually.

2 Summary of steps using VLOOKUP:

- Create a list mapping the terms as they appear in the original data to the new terms to which they are to be transformed
- Use the VLOOKUP function dialogue box to map the original data to the list
- Copy the resulting formula to all the cells to which it is to be applied

B USING THE FUNCTION VLOOK UP

3 Open the RASPBERRY Workbook
Go the worksheet entitled Worksheet 1

- *This represents an extract from a catalogue*

4 Notice that the 'date from' and 'date to' columns (columns E and F) show the date in ISO format (eg 1962-12). This format is useful for sorting, but we wish to transform these dates into a more user-friendly format (December 1962).

5 The 'date from' has already been split into the year and the month components: see columns G and H ('year from' and 'month from number').

- *This was done using the 'Text to Columns' tool on the Data Ribbon, dividing the data at the hyphen: see UNIT G2 for subdividing columns of data*

6 View the Worksheet entitled 'LIST by number'

- *This contains a list of months and their equivalent numbers*

7 Return to Worksheet 1.
Click in cell I4 (in the column headed 'FORMULA month from (name)')
From the function library on the Formulas Ribbon choose 'Lookup and Reference'
From the drop down list, select VLOOKUP

- *A dialogue box appears*

8 In the dialogue box, click in the 'Lookup_Value' box
Then click in cell H4

- *'5' appears to the right of the box, indicating that the value in cell H4 is 5*

9 In the dialogue box, click in the 'table array' box
Then go to the worksheet entitled 'LIST by number'.
Select the whole of cells A1 to B12

- *This tells Excel to use this list as the lookup list.*
- *The 'Table_array' box contains the term: List by number!A1:B12*
- *To the right of the box a summary of the list appears*

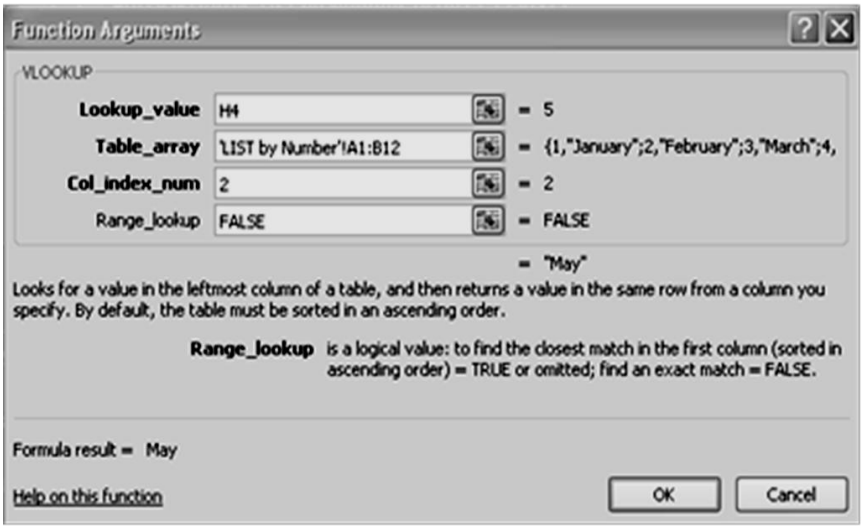
10 In the dialogue box, click in the 'Col_index_num'.
Enter 2

- *This tells Excel to use the value in the second column in the lookup list (column B)*

11 In the dialogue box, click in the 'Range-lookup' box
Type in the work FALSE

- *This is essential to ensure that Excel looks for an exact match (for example, distinguishing correctly between 1 and 11)*

12 Do NOT click OK yet:
Check that the dialogue Box looks as follows:



13 In the dialogue box, 'table_array' box, enter dollar signs before the letters and numbers defining the array: change the term A1:B12 to the term $A$1:$B$12

- *The dollar signs instruct Excel to treat the cell addresses as 'absolute'*
- *Normally when a formula is copied and pasted, the cell address is dynamic and changes to match the appropriate row or columns*
- *In this case, the exact same set of cells must be used throughout.*
- *See UNIT H2 Relative and absolute cell references*

14 Now click OK to close the dialogue box
Observe the contents of cell I4

- *The word 'May' is displayed in cell I4*
- *A formula appears in cell I4 (visible in the formula bar)*

15 Copy the formula to the remaining cells in column I (*month from name FORMULA) use the auto fill handle or copy and paste)*

- *The appropriate month name appears in all cells in I3*
- *Note: if the values do not appear accurate, return to step (13); you may have omitted to transform the cell address to absolute references.*

16 Go to the worksheet entitled 'LIST by number'
Change the name of one of the months (eg to your own name)
Press Enter
Return to Worksheet 1 and observe the effect in column I (*month from name FORMULA)*

- *The new value (month name) is displayed against the appropriate month*
- *The mapping from new to old values can easily be changed without altering the formula.*

C NOTES

17 Remember that the contents of column I are formulas only, not fixed data. To fix the data you will need to copy the columns and paste a new column using Paste Values.

USING THE FUNCTION VLOOKUP
FURTHER PRACTICE

Exercise 1

Open the RASPBERRY Workbook , 'Worksheet 2'
Use VLOOKUP to insert month names for 'month to' in Column M.

- *Date To has already been split into columns (see column L for the month number)*

Exercise 2

Open the RASPBERRY Workbook , 'Worksheet 2'

(a) In column O, use VLOOKUP to transform the month names (in column J) to their French equivalents. Use the list in the worksheet '***LIST by month'***, cells $A$1:$B$12

(b) Once the correct formula has been inserted into column O, go the worksheet 'LIST by month'. Choose one of the other lists (German, Croatian or number) and paste it into column B of 'List by month', overwriting the 'French' list. Observe the results in Worksheet 2, Column O

<< END OF THIS UNIT >>

UNIT K1
UNDERSTANDING LEADING APOSTROPHES AND OTHER SPECIAL CHARACTERS

Purpose:

To explain how Excel interprets special characters such as leading apostrophes

A THEORETICAL BACKGROUND

A1 SPECIAL CHARACTERS

Certain characters have a special meaning when typed as the first character in a cell, and change the interpretation of the rest of the characters in that cell. They are known as 'escape characters'.

They are affected by - or affect - the data type of the cell. (Data Types are covered in the 'essential techniques' course)

This section is mainly theoretical, but if you wish you can open a new, blank worksheet and type in the examples given below as you read through the text.

A2 THE CHARACTERS = + - @

The characters =, +, -, @ all have a special meaning when typed as the first character in a cell (apart from @ they behave as ordinary characters in other positions).

Equals (=) as the first character in a cell defines the cell contents as a formula. If the cell does not contain a valid formula, an error message normally results.

At, plus and minus (@ + and -) are also used within formulas, and using them as the first character in a cell can also generate an error message.

If the cell is formatted as data type Text , then =, + and – can all be used as normal characters.

This does not apply to @. If a cell is formatted as data type text, @ can be used successfully as the first character in a cell. However, even in data type text the character can be interpreted as indicating an email address, and Excel is likely to try formatting any text containing @ as a hyperlink.

A3 APOSTROPHE OR SINGLE QUOTE MARK

Apostrophe, or single quote (‘) at the start of in a cell behaves differently: this character is used in Excel as a shortcut forcing data type Text. There are no exceptions, and the existing data type of the cell makes no difference (even if it is already data type Text).

This only applies to apostrophes at the beginning of a cell (for example: *‘The Book’*). If anything precedes an apostrophe it is treated as normal text (for example: *The title is ‘The Book’*)

Once the apostrophe has been typed, it stops being visible within the cell, though it can still be seen in the Formula Bar. If the document is printed, these ‘invisible’ leading apostrophes do not normally print.

Although the data in a ‘ cell behaves as data type text, using the apostrophe does not cause the number format drop down list to display data type text.

A4 WORKING WITH APOSTROPHES

Whether the ‘disappearing apostrophe’ matters depends on what you are using Excel for. If the Worksheet is for internal use only, and you are creating new data, then getting round the problem may be sufficient.

If possible, avoid using an apostrophe at the start of a cell: if you need to start with quotation marks, use double quotes “ instead. These are treated as ‘normal’ characters. It is also possible to type the starting apostrophe twice: the first one indicates ‘text’ and ‘disappears’ but the second is treated as a normal character.

If leading apostrophes already exist in the data, it is sometimes possible to remove them, but not usually straightforward. They can be replaced manually with double quotes, but not by using search and replace.

Remember that apostrophes are only a problem when they are the first character in a cell. Apostrophes in any other position are treated as a normal character.

A5 PRACTICAL EXAMPLE

The best way to see the effect of these characters is to open a new worksheet and experiment. The following are some suggestions of data which could be entered:

=+-@	Apostrophe
= the book	'Book'
Book = volume	"Book" [double quotes]
+book-volume	''Book'' [2 apostrophes]
-book+volume	The ‘book’
@Book	
Book@volume	

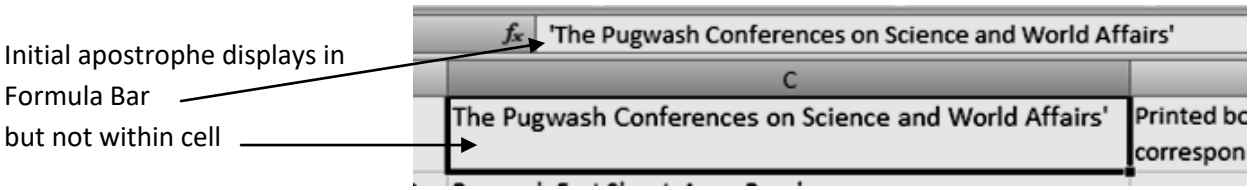
B EXPLORING LEADING APOSTROPHES FURTHER

1 Open the TANGERINE Workbook
- *This is an extract from a catalogue showing a list of publications.*
- *The workbook contains several copies of the list (worksheets Copy 1 to Copy 4)*

2 Click in turn in each of the 'title' cells at item level (cells C3 to C13). For each cell, compare what appears in the CELL with what appears in the FORMULA BAR
- *Some cells (eg C4 and C5) use double quotes (") at the start of the cell: these display as expected.*
- *Some cells (eg C7 and C8) use TWO single apostrophes at the start of the cell. Only the second one displays in the cell (both display in the Formula Bar).*
- *Some cells (eg C10 and C11) use a single apostrophe at the start of the cell, which does not display in the cell (only visible in the Formula Bar).*
- *Some cells (eg C3 and C6) do not have leading apostrophes: apostrophes and quotes within the cell display as expected.*

3 Change three or four of the leading apostrophes, and observe the effect in the cell and in the formula bar. For example:
In cell C4 change the double quotes to single quotes
In cell C7 change the two single apostrophes to one double quote
In cell C10 Insert a second initial apostrophe
In cell C3 Insert single or double apostrophes

4 The leading apostrophe does not display within the cell; you can only see it in the formula bar. This can be irritating if you actually want the text to be in single quotes



5 The feature only applies to the apostrophe at the start of the cell contents. Apostrophes elsewhere in the cell are treated as normal characters.

6 Whether it matters depends on what you are using Excel for. If the Worksheet is for internal use only, and you are creating new data, then getting round the problem may be sufficient. Depending on circumstances, one of the following workarounds might be appropriate:

7 **Option 1:** Avoid using an apostrophe at the start of a cell: if you need to start with quotation marks, use double quotes " instead.

8 **Option 2**: Type the starting apostrophe twice: the first one indicates 'text' but the second is treated as a normal character.

- *If you are printing the document, then the first (leading) apostrophe will not print; if you migrate data to another system, or copy it into Word, then the leading apostrophe is not normally changed to a 'real' apostrophe. In these cases repeating the apostrophe character should not matter. However, this may vary depending on the system you are moving the data into (and its individual settings).*

9 **Option 3:** Enter the data without a starting apostrophe, then add one using CONCATENATE and the function CHAR(39), for example
=CONCATENATE(CHAR(39),E25)

- *Although =CHAR(39) produces what looks like an identical apostrophe, it is not treated by Excel as an escape character.*

10 If the data is to be moved (or printed), experiment with a small selection of the data to see the effect before carrying out any complex procedures to transform it

11 See UNIT K2 *Techniques for removing leading apostrophes* for some suggested methods for removing leading apostrophes in existing data

<< END OF THIS UNIT >>

UNIT K2
TECHNIQUES FOR REMOVING LEADING APOSTROPHES

Purpose: To suggest some techniques for removing leading apostrophes from existing data.

A	***INTRODUCTION*** ➔ *Open the* **TANGERINE Workbook**

1 This unit suggests some techniques for removing leading apostrophes from existing data. It builds on UNIT K1 *Understanding leading apostrophes and other special characters*

2 The leading apostrophe feature is most likely to become a problem with data which has been imported or copied from outside Excel, or which has been entered over time or following inconsistent data entry conventions. In this case, it may be necessary to remove the leading apostrophes.

3 Find and Replace does NOT work on leading apostrophes, nor do data cleaning functions such as CLEAN or TRIM. Changing the data type has no effect.

4 The methods to remove leading apostrophes described below are listed in order of increasing complexity: choose the most appropriate for your data. Although practical examples are given you may choose to omit these on the first reading, and gain an overview of the techniques first.

5 The methods described are as follows
Method 1 Individual editing
Method 2 Copy and paste outside Excel
Method 3 Copy and Paste Values
Method 4 Export as character delimited

6 **Beware:** The results may not always be exactly as predicted below, so always experiment on a small set of data before determining the method to use. The results can be affected by a number of factors, chiefly:
- *The version of Excel, how it has been set up, and the user settings applied*
- *How and in what system the data was originally generated*

B METHOD 1 INDIVIDUAL EDITING

➔ *Open the* **TANGERINE Workbook**

7 Remove each leading apostrophe individually by editing in the formula bar: this may only be appropriate for a small amount of data. You first need to identify which cells are affected.

8 Open – or return to - the TANGERINE Workbook; use a worksheet not previously used for example Worksheet 'Copy 2'
Choose a cell in the 'title' column (column C) which has a single leading apostrophe (eg cell C10): remove the leading apostrophe by editing within the formula bar.
Observe the effect in the cell and in the formula bar.

- *The leading apostrophe is removed*

9 Choose a cell in the 'title' column (column C) which has a two initial apostrophes (eg cell C7): remove the first apostrophe by editing within the formula bar. Observe the effect in the cell and in the formula bar.

- *The leading apostrophe is removed, but the second now becomes the 'leading apostrophe'. To remove leading apostrophes completely BOTH apostrophes must be removed.*

C METHOD 2 COPY AND PASTE OUTSIDE EXCEL

10 Each affected cell is copied individually into Word or a text editor and pasted back in to Excel. This is normally only appropriate for a small amount of data, and you need to identify which cells are affected.

- *The results may differ, depending on how Word has been set up*

11 Open – or return to - the TANGERINE Workbook; use a worksheet not previously used for example Worksheet 'Copy 3'
Choose a cell in the 'title' column (column C) which has a single leading apostrophe (eg cell C11)
Copy and paste into Word
Then copy and paste from Word and paste back into Excel.
Observe the effect in the cell and in the formula bar.

- *The leading apostrophe is removed*

12 Choose a cell in the 'title' column (column C) which has two initial apostrophes (eg cell C8).
Copy and paste into Word
Then copy and paste from Word and paste back into Excel.
Observe the effect in the cell and in the formula bar.

- *The leading apostrophe is removed. The remaining apostrophe is NOT treated by Excel as a leading apostrophe but as a 'normal' apostrophe (because Word has applied its own formatting).*

D METHOD 3 COPY AND PASTE VALUES

13 Copy a whole column (or whole worksheet) and use 'paste values' to paste the data into a new worksheet. This strips out the leading apostrophe, but you will also lose most other formatting.

- *BEWARE: results may be unpredictable, and this normally ONLY works if you paste into cells which have never previously been formatted (changing the data type from Text to General or Number before you paste will not work).*

14 Open – or return to - the TANGERINE Workbook; use a worksheet not previously used for example Worksheet 'Copy 4'
Insert a new column (column D) and format it as Data Type General. Copy the whole of column C. Paste into column D using 'paste values' . Compare each cell with the equivalent in the original worksheet

- *Leading apostrophes have NOT been removed: the 'paste values' method only works when pasting into cells which have never been formatted*

15 Copy the whole of column C. Paste into column A of a NEW, unused worksheet (or even a new workbook), using Paste.

- *The data appears exactly as in TANGERINE Workbook*

16 Copy again and Paste into column B of the same new worksheet, this time using Paste Values. Compare each cell in column A with its equivalent in column B.

- *Leading apostrophes have been removed in column B*

E METHOD 4 EXPORT AS CHARACTER SEPARATED VALUES

17 If none of methods 1 to 3 seem to work, try exporting the data to character Separated Values (CSV) format delimited format and re-import into Excel

- *For character delimited format see UNITS N1 and N2*

F FINAL STEPS

18 **Final Data Cleaning**
Depending on which method you used, you may need to re-apply formatting, layout and column widths.

- *Beware: copying and pasting formats (or using the format painter) will re-insert the leading apostrophes.*

19 **Re-applying formatting**

In the new worksheet used in the steps 15 and 16 above, transfer the formatting from column A to column B using 'format painter' .

(Alternative: copy column A and paste the formats only into column B using Paste Special> Paste Formats).

Compare each cell in column A with its equivalent in column B.

- *Leading apostrophes have returned in column B, demonstrating that Excel treats them as an aspect of formatting, not as text characters*

20 Once the leading apostrophes have been removed by one of the methods described above you may be left with data missing a 'genuine' opening apostrophe, or some cells with and some without an opening apostrophe, or a mixture of double and single apostrophes. These can be addressed with a combination of concatenation and Find and Replace.

For example:

Starting data	'Quotation'
Leading Apostrophe removed	Quotation'
Formula to insert double quote at the start [Note that this needs FOUR sets of double quotes]	=CONCATENATE("""",C4)
Result of formula	"Quotation'
Find and replace to change ' to "	"Quotation"

21 Beware: if you use concatenation or find and replace to insert a leading apostrophe, you will be back where you started, with an 'invisible' leading apostrophe (though the data will at least be consistent).

22 Alternative: use CONCATENATE with the function CHAR(39) to add a single quote mark **=CONCATENATE(CHAR(39),E25).**

23 **Excel Options**

If you are still having problems after trying all the above methods, it is possible that one of the Excel Options has been changed. In **Excel Options > Advanced Options** ensure that the **'**Ensure Transition Navigation Keys' option is NOT ticked. If it is ticked ALL cells will have a leading apostrophe inserted.

VERY IMPORTANT: IF YOU EXPERIMENT WITH CHANGING THIS OPTION, ENSURE THE BOX IS UNTICKED BEFORE YOU MOVE ON TO A NEW EXERCISE OR ACTIVITY.

REMOVING LEADING APOSTROPHES
FURTHER PRACTICE

Exercise 1

Using any of the data in the TANGERINE Workbook (or data you have created yourself) remove the leading apostrophes using any of the suggested methods: which works best?

Exercise 2

2 If you know how to carry out data imports and exports using **Character Separated Values** format (covered in units N1 and N2) try this method to remove leading apostrophes

Exercise 3

3 Study any of the data used in this Unit (the original data in the TANGERINE Workbook, data from which the leading apostrophes have been stripped, or data which you created yourself). Consider how best to ensure data consistency, using a combination of CONCATENATE and Find and Replace. If you have time, try carrying out the procedures and see their effect.

<< END OF THIS UNIT >>

Excel for Archivists Workshop

Data Improvement and Data Migration

Part 2 Data Migration
(Units M - P)

HANDBOOK

Gillian Sheldrick
2022

Version 1.7

[blank page]

UNIT M1
INTRODUCTION TO DATA MIGRATION

This UNIT covers the theory only; there are no practical examples

A What is Data Migration?

1 **Data Migration** means moving data from one system to another. In practice, data is almost always COPIED not moved; a copy of the data normally remains in the original location.

2 During a migration, data is exported from the system or format in which it originated and imported into somewhere new. However the terms 'import' and 'export' can obscure what is actually happening and where the control lies.

3 Data may be:
- '**Pushed out'** of one system or format into another
- **'Pulled into'** a new system or format from an old one

(IT developers often talk about 'sucking the data in' or 'spitting it out')

4 Sometimes it is technically easier to move data to an intermediate system or format rather than to move it straight from the source system to the destination. A **Data Exchange Format** is often used for this purpose. Excel can 'push' data out into a few standard data exchange formats (see UNITS N1 to N4 for more on data exchange formats).

5 In order to ensure the success of a particular migration it is essential to understand where the control lies (is the data being 'pushed out' or 'pulled in'?) and whether a data exchange format or interim system is required.

6 Whether pushing or pulling, the data being moved from one system must be **compatible** with the requirements and structure of the destination system (UNITS M2 and M3 deal with data compatibility)

B The role played by Excel and other systems

7 What Excel can and cannot do:
- Excel can 'push' data out into a few standard data exchange formats, principally Character Separated Values (CSV) and XML (see UNITS N1 to N4)
- Excel can also 'pull data in' from these formats
- Excel on its own can neither 'pull data in' from, nor 'push it out' to Archives cataloguing systems such as Calm, Adlib, the Archives Hub etc

8 The role of other systems:
- Most archives catalogue systems and other database systems have built-in tools which enable them to 'pull data in' either direct from Excel or using a data exchange format.

THE WATER PISTOL ANALOGY

Pushing data out

- The water pistol squirts water out, but if it contains a mixture of pink and blue drops you can't choose which are expelled – the water will come out purple.
- Similarly, pushing data out of a system may not be very precise (for example, you may not be able to choose exactly which fields are exported)

Pulling data in

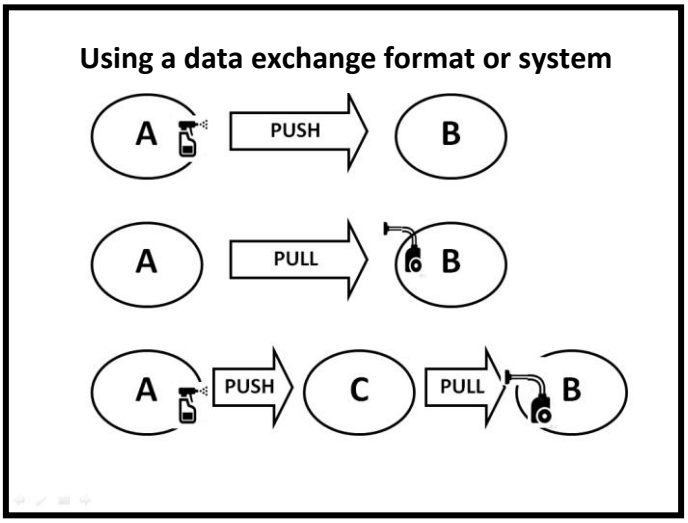
- The water pistol can be selective about which dish (pink or blue water) it sucks from
- Similarly, pulling data into a system often gives more control (for example, choosing which fields and which records are to be imported)

Data compatibility

- The water pistol is unlikely to work using sand instead of water (whether you are trying to squirt it out or suck it in)
- If data is incompatible, it is normally impossible to move it successfully into another system

Using a Data Exchange Format

- The water pistol cannot suck up – or squirt out – large crystals of rock salt. But it can if you first dissolve the salt in a glass of water.
- A similar intermediate stage - a data exchange format - might be needed to enable a data migration to work



<< END OF THIS UNIT >>

UNIT M2
DATA COMPATIBILITY AND DATA STRUCTURE

This UNIT covers the theory only; there are no practical examples

1 What must be compatible?

Data can only be transferred successfully from one system to another if everything is compatible. This includes:

- Hardware
- Medium and format
- Data structure (in columns, in rows, within each cell, and representing the archival hierarchy)

2 Hardware compatibility

This is fairly obvious, for example:

- If the data is planned to be transferred on CD, the computers at both ends need a CD drive.
- If it is to be transferred online, an internet connection is required at both ends

3 Medium and format

This is about the relationship between Excel and the system the data is being moved to. For example:

- Can the recipient system accept data direct in Excel format (as a *.xslx file)?
- Must the data be moved first into an intermediate format, for example as a Word (*.docx) or Text (*.txt) file
- Must the data be moved first into a specific file transfer format *[see units N1 to N4]* , for example in Character Separated Values (CSV) or XML or EAD?

4 Data Structure and Compatibility

The way the data is structured within Excel must be compatible with the requirements of the destination system. There are three main aspects to this:

- Across: the data elements in Excel (the columns) must replicate the field structure and table structure, or the element structure of the recipient system.
- Within each cell: the data must match the rules applied to the equivalent element or field in the recipient system.
- Down: the data must be divided into rows in a way which corresponds to the data requirements of the recipient system, especially the way the archival hierarchy is expressed.

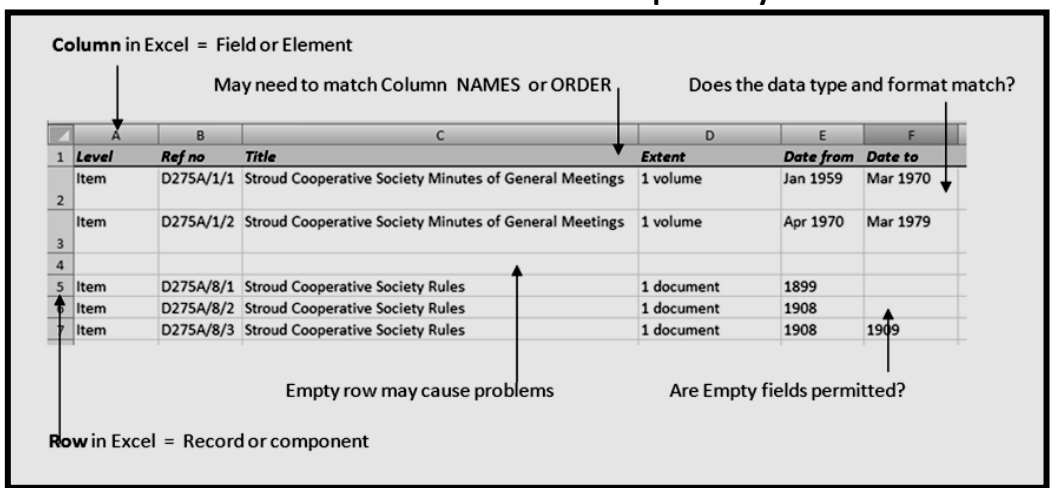
5 Data Structure and Compatibility: Columns

- Each COLUMN in Excel is the equivalent of one database FIELD or one ELEMENT (for example, reference number or title)
- Depending on the destination system, it may be essential to give each column an identical NAME to the name of the field or element as in the destination system
- It may be essential to have the columns in the same ORDER as the fields or elements in the destination system
- It may be essential to have the same NUMBER of columns as the fields or elements in the destination system (in which case some columns may contain no data)
- One type of data which often varies between systems is the way date data is structured: for example, dates from and dates to in a single column or in separate columns; months and years in the same column or in separate columns.
- If the data is destined to be transferred to multiple database TABLES, it may be simplest to migrate into each table independently, and make the links between them in the destination system rather than in Excel. Replicating the structure of multiple database TABLES in Excel is possible, but not straightforward; it is beyond the scope of the Workshop.

6 Data Structure and Compatibility: Rows

- One row in Excel normally represents one record or component in the destination system. Spreading data for a single record across more than one row is likely to prevent the migration working as expected
- Leaving rows empty (for example between different sections of the catalogue) may prevent the migration working

Data Structure and Compatibility



	A	B	C	D	E	F
1	***Level***	***Ref no***	***Title***	***Extent***	***Date from***	***Date to***
2	Item	D275A/1/1	Stroud Cooperative Society Minutes of General Meetings	1 volume	Jan 1959	Mar 1970
3	Item	D275A/1/2	Stroud Cooperative Society Minutes of General Meetings	1 volume	Apr 1970	Mar 1979
4						
5	Item	D275A/8/1	Stroud Cooperative Society Rules	1 document	1899	
6	Item	D275A/8/2	Stroud Cooperative Society Rules	1 document	1908	
7	Item	D275A/8/3	Stroud Cooperative Society Rules	1 document	1908	1909

7 Data Structure and Compatibility: Within each cell

Two types of requirements need to be considered: technical and stylistic.

If TECHNICAL requirements are not adhered to the migration will probably fail to work. For example:

- Data may need to be formatted as a particular data type (eg date or text or number).
- Certain characters may be forbidden or compulsory (for example, non-printing characters such as line ends may need to be avoided; certain types of data may need to be in quotation marks).
- There may be limits to the number of characters in a field or element).
- Cells containing no data may not be permitted.

Failing to meet STYLISTIC requirements, which are normally in-house preferences and conventions, will not normally affect the migration. However, it is often easier to apply the desired style using Excel than in the destination system. For example:

- names of months may be in full or abbreviated;
- certain data may be in upper case;
- there may be preferred ways of expressing personal names.

8 Representing the archival hierarchy

Care must be taken to represent the archival hierarchy in Excel in a way compatible with the requirements of the destination system. This includes using the appropriate column structure, the correct order of rows and the correct data within cells. This topic is dealt with separately in UNIT M3.

<< END OF THIS UNIT >>

UNIT M3
REPRESENTING THE ARCHIVAL HIERARCHY

This UNIT mainly covers the theory, but includes some practical examples

Excel Workbook: This unit uses the RAISIN and PRUNE workbooks

A Introduction

1 The archival hierarchy and levels of description (fonds – series – item etc) can be represented in a variety of ways in Excel, in databases, and in other archive systems. In order to migrate data successfully, the methods used in the source (Excel) and the destination must be compatible (see Unit M2 for more on compatibility).

2 There are two possible approaches when migrating data from Excel to another (destination) system, described in sections B and C below.

- Approach 1: Apply the hierarchical structure AFTER migration, using the functionality of the destination system
- Approach 2: Structure the data in Excel BEFORE migration in a way compatible with that used in the destination system

B Approach 1: structure the data AFTER migration

3 This may be appropriate if:

- The destination system offers a simple and effective way of making and changing links between levels, and adding additional levels
- The archival structure is relatively simple (eg many item level records and few higher level records)
- The total number of records is small

4 Method: In Excel

- Create the catalogue in a logical order with one record per Excel row.
- 'Pretend' that the list has a completely flat structure, with every record at the same level of description (eg Item)
- Keep a back-up copy indicating the genuine levels of description

5 In the destination system:

- Depending on your system, you may need to create a fonds level record before migration, and then migrate all the other records as 'item' level records.
- Once the data has been migrated, use the functionality of the destination system to structure the data into the correct archival hierarchy

6 For an example of a list prepared structured ready for export as a flat list, open the RAISIN Workbook, worksheet 'flat list'. In this example the data will be migrated as if all records below fonds level are item level; the correct links and levels will be applied after migration.

C Approach 2: Structure the data in Excel BEFORE migration

IMPORTANT:
The methods described below are generic; the exact method used must match that prescribed by the rules for the system you are migrating into

7 This may be appropriate if:

- It is complex to make links between levels in the destination system
- The archival structure is complex (eg many different levels of description)
- The total number of records is large

8 With this approach, it is essential to understand the methods used in the destination system to link records at different levels of description, and to reflect this in the Excel data.

9 There are three basic ways used to link records into a hierarchical structure. Sections D, E and F below describe for each a conceptual model illustrating the key features, followed by the method used to structure the data in Excel.

10 Note that in each case the description and examples are generic. In practice, each system will have its specific requirements (for example, you may need to follow system rules for naming the columns or structuring the data)

11 *Nickname for conceptual model:* CHAINS
Technical name for the method: parent and child links
Described further: Section D

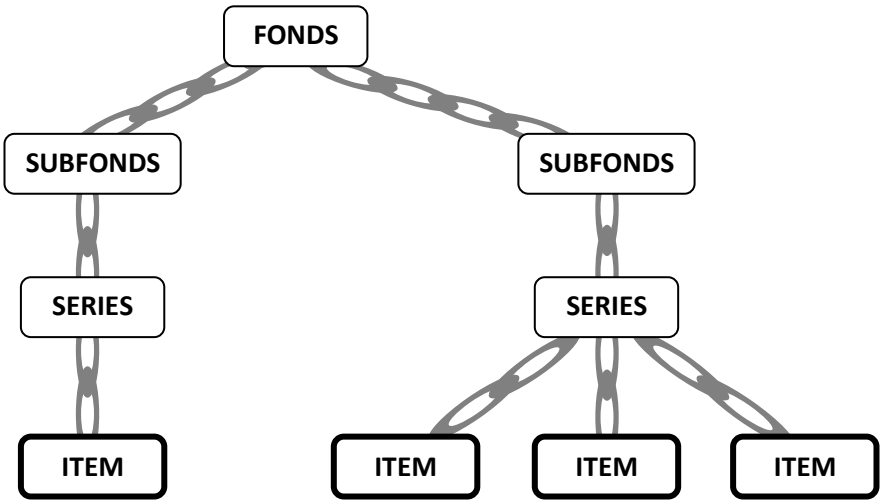
12 *Nickname for conceptual model:* RIBBONS
Technical name for the method: parent and child links [a more rigid variation of the method]
Described further: Section E

13 *Nickname for conceptual model:* BOXES
Technical name for the method: nested components
Described further: Section F

D Structuring data to reflect the archival hierarchy: Method 1 'CHAINS'

Technical term: Parent-child links

Conceptual model: Each record is linked to its immediate parent, forming a series of chains; for example, each item level record is linked to its parent series and each series to its parent subfonds.



Effect of moving and inserting records:

- To insert a new item level record, a single chain needs to be added, to link it to its immediate parent
- To move a subfonds to a different fonds, only the subfonds link has to be broken and re-attached; all the records hanging off it remain chained together.

Examples of systems using this method: Adlib; most systems based on Access or other traditional databases.

Reflecting the structure in Excel

- Records can be listed in natural order, with each higher level record followed in turn by each lower level record. However, this is not always essential, and other orders are theoretically possible (for example, all series level records could be listed first)
- Each record is assigned a reference which is unique within the whole database, often referred to as a unique identifier (UID).
- UIDs can in theory be in any format; they bear no meaning in themselves, and can be compared to bar codes used to identify boxes.
- Conventionally a sequence of running numbers (eg 673458) is used (though UIDs could in theory appear in any order so long as they are unique within the database)
- Each record contains a field to record the unique reference of its immediate parent. For example, item level record 673458 may be a child of (linked to) series level record 673399.

Reflecting the structure in destination systems:

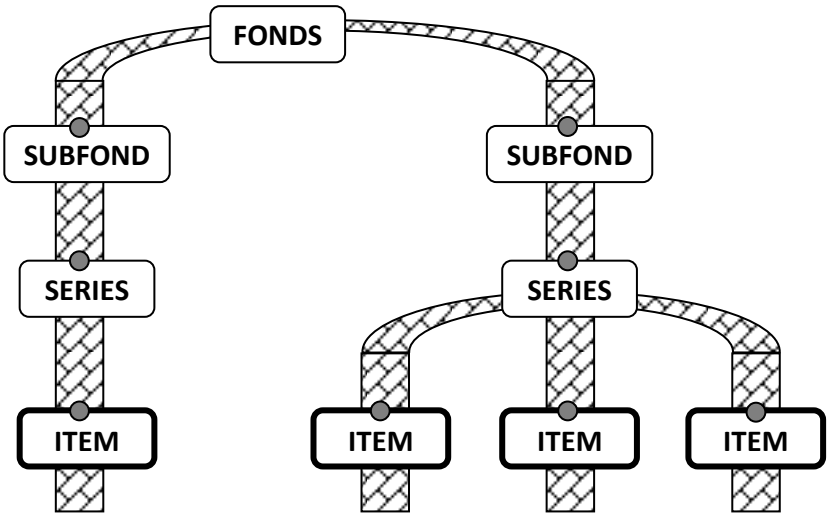
- As described above, each record is assigned a unique reference and includes a field to record the unique reference of its immediate parent.
- Before migrating the data, it will probably be necessary to amend the Unique identifiers and parent records to match the sequence used in the destination system and ensure that they do not duplicate values already in the destination system
- This can be done in Excel using a formula (for example, to add 599999 to each UID and parent UID)

Example in Excel: Open the **RAISIN Workbook**, worksheet 'Chains'. This includes fields for 'Unique number and 'parent number'.

E Structuring data to reflect the archival hierarchy: Method 2 'RIBBONS'

Technical term: [this is a variation on the parent-child link]

Conceptual model: The outline structure consists of a series of ribbons 'hanging' from the fonds level record. Individual records are 'pinned' to the appropriate place on the ribbon structure. *[Note that the analogy for this method is less exact than for 'boxes' or 'chains'].*



Effect of moving and inserting records:

- To insert a new item level record, a new ribbon is added in the appropriate place
- If a subfonds is moved by unpinning the record, the children hanging below it remain pinned to the ribbon, with no parent subfonds; to move them each has to be 'unpinned' individually.

Examples of systems using this method: used by Calm

Reflecting the structure in Excel – and in Calm

- Records are listed in natural order, with each higher level record followed in turn by each lower level record.
- The reference number is used to link parent and child records together: the reference number for the child record is formed by adding a sub-number to its parent record
- For example, record D275/1/2 is by definition a child record of D275/1. If the reference number of either is changed, the link is broken.

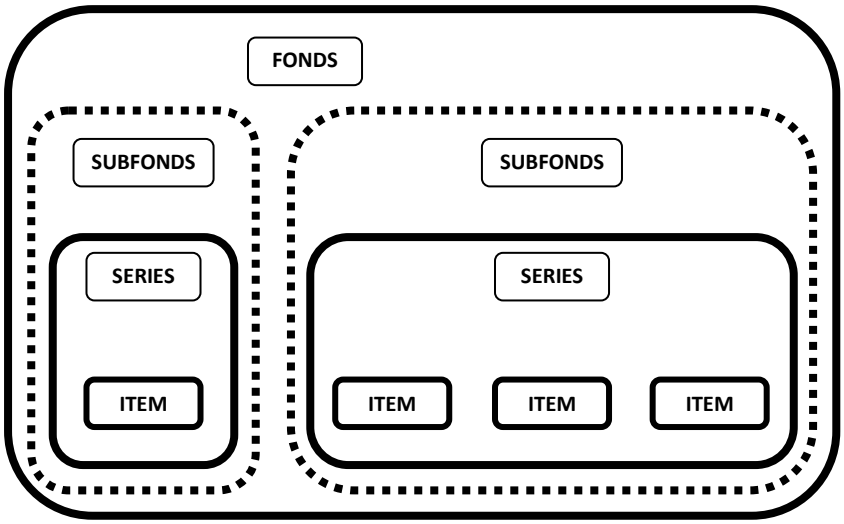
If the preferred reference numbers are not in the correct format, for example D275-1A/3 or D275-1A(4), these have to be recorded as an 'alternative identifier'.

Example in Excel:
Open the **RAISIN Workbook**, worksheet 'Ribbons'. The column 'structured reference number' links the records. The column 'Alt Ref No' contains the 'real' reference number, which is in alphanumeric format, so cannot be used to link records.

F Structuring data to reflect the archival hierarchy: Method 3 'BOXES'

Technical term: Nested components

Conceptual model: The item level records for each series are enclosed within a box; the series 'boxes' for each subfonds (with their enclosed item level records) are enclosed within a larger box; the subfonds 'boxes' for each fonds (with their contents) are enclosed within a box (like a set of Russian dolls)



Effect of moving and inserting records:

- To insert a new item level record, all the 'boxes' must be 'opened'
- To move a subfonds to a different fonds, only the subfonds 'box' needs to be moved; the entire contents of this 'box' move with it

Examples of systems using this method: XML; EAD (Encoded Archival Description); the Archives Hub.

Reflecting the structure in Excel

- Records are listed in natural order, with each higher level record followed in turn by each lower level record.
- Additional columns are inserted to indicate the start and end of each component.

Reflecting the structure in other systems: In XML the 'top and bottom' of each 'box' represent the start and end of each component, as defined by a pair of tags (for example <c01> and </c01>). Like the boxes, the components are nested within each other. For more on XML see UNIT N3

Example in Excel: Open the **RAISIN Workbook**, worksheet 'Boxes'.

- The START of each component ('box lid') is indicated in column B
- The END of each component ('box base') is indicated in columns I to L: Note that some rows (eg row 7) contain no component end while in others (eg rows 13 or 29) there are two or more component ends
- An additional row has also been added at the end of each component for clarity: these rows must be removed before migrating the data.

REPRESENTING THE ARCHIVAL HIERARCHY
MORE PRACTICE

EXERCISE 1

Open the **RAISIN Workbook**

Study each worksheet as follows, and compare them to ensure you understand how each method works and how it differs from each of the others:

- Sample catalogue (the original, unamended data)
- Flat list (hierarchical structure to be applied following migration)*[UNIT M3 section B]*
- Chains *[UNIT M3 section D]*
- Ribbons *[UNIT M3 section E]*
- Boxes *[UNIT M3 section F]*

EXERCISE 2

Open the **PRUNE Workbook**
For each of the following worksheets in the PRUNE Workbook (copies of those in RAISIN), amend the data as appropriate so that the whole series **D275-1D *Stroud Cooperative Society: Superannuation Committee Minutes***, including its three items becomes a series within sub-fonds D275-2

- Chains
- Ribbons
- Boxes

Hint: follow the correct rules for each of the three methods. This may include moving rows and/or changing the data

EXERCISE 3

Open the **PRUNE Workbook**

For each of the following worksheets, identify which method has been used to express the hierarchy ('boxes' 'chains' or 'ribbons'). For answers and comments see the sheet labelled 'Answers'.

- Example 1
- Example 2
- Example 3
- Example 4

<< END OF THIS UNIT >>

UNIT M4
PLANNING A DATA MIGRATION PROJECT

This UNIT covers the theory only; there are no practical exercises

A Planning

1 Allow enough time: the procedure can be complex, and you may need more than one attempt.

2 Migrate responsibly: it is easy to destroy or corrupt existing data if you don't understand what you are doing

3 You must understand:
- Your data
- Your source system (eg Excel)
- Your destination system
- Any data exchange format or system
- The overall environment
- The relationship between all these elements

4 Researching the migration procedure and requirements in your specific environment is as much part of a cataloguing project involving migration as researching the administrative history or archival structure.

5 Understanding the data:
- Is the data and its structure compatible? (see UNIT M2)
- How does the existing data relate to the destination system (see UNIT M5)
- How is the archival hierarchy controlled? (see UNIT M3)
- Are all mandatory fields populated?
- Is the new data is being added at the end of the existing data (appended) or replacing (overwriting) existing data

6 Understanding the systems:
- Have you had appropriate training?
- Does your system have a specialist data import tool?
- If so, how is it accessed?
- Do you need a particular administrator permission or login?
- How do you select which fields will be imported?
- What are the system names for the fields (not necessarily the same as the field labels you may be used to)?
- How will you undo a migration which hasn't worked as expected?

7 Understanding the environment:
- Can you carry out the migration yourself or do IT specialists need to be involved?
- Is the data being pushed out of the source system or pulled into the destination system? (see UNIT M1)
- Is a data exchange format or system required? (see UNITS M1; N1 – N4)
- Is a test system available?

8 Map the data in your source system to the requirements of the destination system (see UNIT M5 data Mapping)

B Preparation

9 Before you start, take backups:
- Of the source data (eg the Excel workbook)
- Of the destination database

10 If you have a test system, use it for a trial run first

11 Do a trial run on a small sub-set of the data first

12 Ensure the data is correct and complete (it is normally easier to check and amend in Excel than after the migration)

13 Ensure the data structure and content is compatible with the destination system (UNIT M2) including:
- Across (the columns)
- Within each cell
- Down (rows)

14 In Excel, remove
- all filters
- all merged fields
- any blank rows

15 In Excel, remove any data not required to migrate. This is especially important if it does not conform to the required format (for example, headings or comment rows inserted within the worksheet)

16 Ensure that any values which must be unique (for example, reference numbers) are not duplicated between the Excel data and the destination database

17 Make a note of any manual alterations to be made in the destination system following migration

C After Migration

18 Don't assume that the migration has worked accurately: always check the data

19 Have all the records have been imported?

- This may be as simple as making a note of the number of records in the database before and after the migration and comparing with the number of Excel rows you intended to import (remember that row 1, the header row, is not normally imported)

20 Check a sample of records to ensure that all the fields have been imported and the data looks as expected.

21 Check the records are in the correct position in the archival hierarchy and that they are attached to the correct parent.

22 Do the rules in your data map (see UNIT M5) need amending or refining?

23 Check that the data already in the system hasn't been corrupted or overwritten.

24 Don't forget to make any manual amendments to the data

D What went wrong?

25 Reasons for import failure can include the following:

- No data in mandatory fields
- Repeating a value (such as a reference number) which is supposed to be unique
- The wrong type of data in controlled fields (for example, a medium of 'photograph' may be allowed, but not 'photo')
- Trying to migrate characters which the database cannot recognise.
- Migrating a formula from Excel (instead of the values)

<< END OF THIS UNIT >>

UNIT M5
DATA MAPPING

This Unit is in two parts:
PART ONE (sections A, B and C) introduces the concept, including a simple example
PART TWO provides a practical exercise using the QUINCE workbook

UNIT M5 PART ONE

This part introduces the concept and theory of Data Mapping

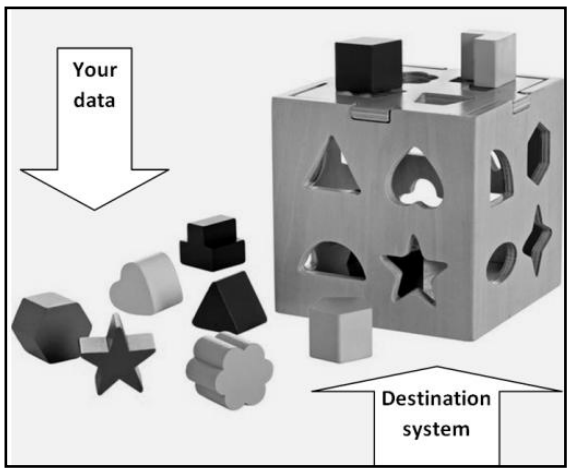
A Introduction

1 *What is Data Mapping?*

As part of planning a data migration, you will usually need to compare the data you already have with the requirements of the system it will be moved to. This might be as simple as matching (or mapping) existing column headings in Excel to the elements available in the new system. However, it is also likely to involve defining rules for transforming your data in some way.

2 *An Analogy*

Think of the destination system as a shape sorter, into which the 'shapes' of your data must be slotted.
Each shape (each element of your data, for example 'reference number') must be matched with the appropriate 'hole' in the destination system.
Some shapes (elements) will fit exactly. Some may need to be reshaped slightly to fit into a hole. Some shapes (or holes) may be missing altogether and might have to be created.



3 *Who is the Data Map for?*

At the start of the migration project, it will be primarily for yourself, as a tool to help you understand your own data and the system into which it will be moved. Later on, it may form part of a specification for consultants or IT developers building migration tools. If you are the sole audience, you may not need to include much detail; if you are using it to brief others, you may need to define rules very precisely.

B Scope

4 ***What does a Data Map contain?***

Depending on its audience, a data map is likely to contain some or all of the following:

(1) A list of data elements, fields or columns available in the destination system, matched to their equivalents in your source data. For example, the column you call 'Reference Number' might contain the same type of data as the destination system's 'Identity Code' element.
(2) A list of general requirements for the destination system, with a description of how your source data meets or does not meet those requirements. For example, the columns of data may appear in a different order in your source data and in the destination system.
(3) Rules for transforming your source data, either to meet the requirements of the destination system, or to improve the data or make it more consistent before carrying out the migration. For example, 'Remove full stop at the end of the title field'; 'Combine data from arrangements and accruals column'. Developing and refining these rules is likely to be the most time consuming aspect of the data mapping process.

5 ***Additional contents***

If you are intending to carry out the work yourself using Excel, you may also choose to include notes on which Excel techniques you might use to implement the data transformation rules you have defined. However, this is not strictly part of data mapping. If someone else (especially a professional IT developer) will be carrying out the work, you should not try to define HOW the rules should be implemented, but focus on defining clearly WHAT you expect to achieve.

See overleaf for an example of a data map

6 *Example*

The following very simple example is based on the tasks carried out in units G2 (Subdividing columns of data) and G3 (Removing duplicate records), using the BLUEBERRY and CHERRY workbooks. It is intended only as an illustration of the concept, not as an example of best practice.

Example of source data (extracts from BLUEBERRY workbook)

Parish	*Reference*	*Name and status*	*Year*	*Type*
Longbridge Deverill, Wiltshire	*P2/3Reg/182D*	*Waters, Thomas, Husbandman*	*1559*	*Will*
Crockerton, Longbridge Deverill, Wiltshire	*P2/4Reg/234B*	*Adlam, Joan, Widow*	*1566*	*Will*

Example of data mapping (individual elements and general requirements)

DATA ELEMENT (DESTINATION SYSTEM)	*MAPPED DATA ELEMENT (SOURCE DATA)*	*NOTES AND RULES (INDIVIDUAL ELEMENTS)*
Reference	Reference	No changes required
Date	Year	Change column heading to 'Date'; no other changes required
Name and status	Name and status	No changes required
Place	Parish	(1) create new column headed 'Place' (2) 'Place' is all the data before the comma. If there is more than one comma, it is all data preceding the final comma.
County	Parish	(1) create new column headed 'County' (2) 'County' is all the data after the comma. If there is more than one comma, it is all data following the final comma.
Document Type	Type	Change column heading to 'Document Type'; no other changes required

GENERAL REQUIREMENTS	*RULES FOR IMPLEMENTATION*
Rows must not be duplicated	(1) Interpret 'duplicate' as follows: if 'Reference' and 'Name and status' are both identical, interpret rows as 'duplicates', even if data in other fields differs (2) Remove rows which are 'duplicated' according to the above definition.
Column headings must match those in destination system and must appear in the identical order	(1) changes to column names noted in notes and rules for individual elements (2) change column order to appear the following order left to right: Reference, Date, Name and status, Place, County, Document Type

C THE DATA MAPPING PROCESS

7 *Where is the Information I need?*

The are likely to be three equally valuable sources of information, for both the source data and the destination system:

- Examination of the data
- Formal rules for the system
- System experts

8 *Examination of the data*

- **Source data:** Although you may be familiar with the system as currently used, if data has been created over many years the standards used may well have changed. For example, perhaps Parish and County were initially recorded in a single column, with a separate column for County being added at a later date. A useful strategy can be to start with a very small extract of the data, say ten rows created at different dates. Then after the first attempt at mapping, take a larger sample and see if the rules you have developed need amending. You will probably need to continue this process iteratively, even after the first migration test has been carried out.

- **Destination System**: If data in the destination system is derived from multiple sources (eg Archives Hub) it may be worth asking the system owners to identify some appropriate data which is similar to yours and represents good practice. Randomly-chosen data might meet all the technical requirements but use a completely different house style. You may decide to include rules in your mapping to make your own data more consistent with existing data. Study what your own data looks like in the new system following each test migration and modify your mapping accordingly.

9 *Formal rules for the system*

- **Source data**: if the data is currently in a proprietary system (eg Calm), study the system manual. If the data is in a simpler system, such as Excel, formal documentation may not exist.

- **Destination System**: As you are likely to be less familiar with the destination system, it is particularly important to study the documentation and rules, both for the system as a whole and the rules for specific elements / fields.

10 *System experts*

- **Source data:** You may yourself be the expert in your own system and data. If not, try to track down the people who wrote it in the first place. Don't forget technical experts involved in maintaining the system. People who use your current system on a day to day basis, though not necessarily experts in the 'system', are likely to have a good knowledge of the data and how they record it; they may well have developed their own unrecorded 'rules'.

- **Destination System:** An existing system should be supported by a network of experts on different aspects, in additional to any written documentation. Make use of as many of the following as you can: Help desk; training courses; user group and other colleagues; technical expertise.

11 *What Questions Shall I Ask?*

The following suggestions provide a starting point, but many more questions are likely to arise as you develop and test the mapping. Your first attempt at mapping is very likely to include notes of issues which cannot immediately be resolved.

- Must the columns (fields) be in a specific order?
- Must the column headings have specific names?
- May cells be left empty (ie is any data mandatory)?
- May data be duplicated (ie are unique values mandatory, say for reference number)?
- Will omission of mandatory data prevent data migration working? (The migration routine might interpret a blank field as indicating the final record and omit all subsequent records, or might simply stop when it encounters a data 'error'.
- Sometimes though, 'mandatory' just means that the data is expected to conform to the rule, but 'incorrect' data will still migrate.
- How are levels of archival description - and the links between them - expressed?
- What is the maximum number of characters permitted in each cell?
- How must the data in each cell be structured (for example, must specific data types - number formats - be used)?
- What type of data must be entered in each cell (for example, dates from and dates to in a single column or in separate columns)?
- Are certain terms preferred or banned (eg 'photograph' may be permitted in the extents field but not 'photo')
- Must non-printing or other characters be removed before migration (for example, inclusion of hidden line ends or initial apostrophes may cause technical problems or unexpected results)?
- Are there limits to the number of characters in a field or element? This may not be documented, and may not become apparent until you test the migration

UNIT M5 PART TWO

This part contains a practical exercise, using data in the QUINCE workbook

BACKGROUND AND OPTIONS

In this exercise you map the data provided in the QUINCE workbook to the requirements of a fictional destination systems ('Archisearch'). The assumption is that data will eventually be migrated from QUINCE to Archisearch, but this exercise does not include carrying out any migration. The exercise does not include any consideration of how to express the archival hierarchy (for which see unit M3).

There are **two options** for carrying out the exercise (more detailed instructions for each option are given below):

- Option 1: Study the data and the rules for the destination system alongside the suggested answers (this option is recommended if time is limited)
- Option 2: Study the data and the rules for the destination system, and attempt to carry out some the mapping yourself before looking at the suggested answers.

OPTION 1

Study the data and the rules for the destination system alongside the suggested answers

OPTION 1: TASK 1

- Open the QUINCE workbook which contains samples of the source data, and familiarise yourself with the structure of the data.
- Turn to Table A below, rules A1 – A8. This sets out the general rules which will apply to the data migration. Read the rules and the data mapping notes, referring to the data in QUINCE if you wish. You do not need to add anything to this table.

OPTION 1: TASK 2

- Turn to the table 'SUGGESTED ANSWERS FOR TABLE B' (Unit M5 pages 11-13 below)**.** This sets out the rules applicable to each individual element available in 'Archisearch', together with the equivalent elements in the data sample (QUINCE), data transformations which might be required in order to confirm to the requirements of Archisearch, and notes on how this might be achieved using Excel.
- Read the notes for each element (or choose just two or three to study) and ensure you understand why the 'suggested answers' are applicable.

OPTION 2

Carry out some the mapping yourself

OPTION 2: TASK 1

- Open the QUINCE workbook which contains samples of the source data, and familiarise yourself with the structure of the data.
- Turn to Table A below, rules A1 – A8. This sets out the general rules which will apply to the data migration. Read the rules and the data mapping notes, referring to the data in QUINCE if you wish. You do not need to add anything to this table.
- Turn to table B below, rules B1 – B9 and X1 – X3. This sets out the rules applicable to each individual element available in 'Archisearch'.
- For each element, use the 'your notes' area to write the name of the column in the source data (QUINCE workbook) which maps to the 'Archisearch' element name. For example, 'Description Level' in the QUINCE workbook maps to 'Level of Archival Description' in Archisearch.
- If you wish, refer to the suggested answers (Unit M5 pages 11 - 13 below).

OPTION 2: TASK 2

- For each element (or choose just two or three elements) consider what special rules might be required in order to conform to the requirements of Archisearch. For example, it may be necessary to change data into upper case, or combine data from two Excel columns.
- The 'Hints' in the left-hand column of the QUINCE Workbook point to some (not all) possible anomalies which might need to be addressed before or during the migration process.
- If you wish, consider which Excel techniques might be used to transform the data according to the 'Data Mapping Rules' you have defined. This is not strictly part of data mapping
- As this is a planning exercise only, you do NOT need to make any changes to the data in the QUINCE workbook (unless you wish to).
- You can either make brief notes or refer to the suggested answers (Unit M5 pages 11 - 13 below)

DATA MAPPING EXERCISE

Table A General Rules

	ARCHISEARCH Rule	***Data Mapping Notes (using data sample in QUINCE Workbook)***	***Notes on using Excel to implement data changes***
A1	Data must be submitted in Excel (any version)	*No changes required – already in Excel*	*Not applicable*
A2	Only the elements (Excel columns) listed in the rules for individual elements (below, table B) may be used; no additional elements may be added	*Need to map old data to new elements.* *The existing data includes the following elements but there is no equivalent in Archisearch. See notes in element mapping (X1 to X3) for details:* • Accession number • Date details • Box / location	*Not applicable*
A3	For each element (column), the data must conform to the rules detailed below	*See notes for each element*	*Not applicable (see individual field notes)*
A4	Data must be entered in columns listed as mandatory: cells in these columns may not be left empty	*The following element is mandatory in Archisearch but there is no equivalent in existing data. See notes in elements mapping for details:* • Access restriction	*Not applicable (see individual field notes)*
A5	Column names must match those used in Archisearch (as given in rules below)	*The following Archisearch elements have different names in the existing system See notes in element mapping for details:* Scope and Content ReferenceCode Year from Year to Extent/Quantity	*Enter new names manually*
A6	Columns must appear in the same order, left to right, as the order in which the Archisearch fields are listed below	*Rearrange elements to match order listed in table B (from left to right)*	*Cut and paste*

A7	No rows may be left empty	*Remove blank rows before 'series' and 'section' level rows*	*Can use 'remove duplicates' (but this will leave a single blank row which is not 'duplicate': remove this manually).*
A8	Data must not contain hidden line ends or other non printing characters	*Check for and remove hidden characters.* ***NOTE: in this sample there is only one instance. D7338/10/7 'Description' has hidden line ends. These characters are visible if you copy the cell and paste into Word.***	*The function CLEAN will remove hidden, nonprinting characters. Excel does not have a function which will identify such characters without removing them. It is possible to do this by using the 'LEN' (length) function before and after applying 'CLEAN', to identify which cells are shorter as a result of removing characters, but this is a bit laborious.*

Table B Rules for Individual Elements

(blank copy if you wish to try developing rules yourself: suggested answers below)

	ARCHISEARCH Element Name	***ARCHISEARCH Element Rules***	***Your Notes***
B1	Level of Archival Description	Mandatory. The following terms are permitted: Fonds; Series; file; item (together with 'sub-levels such as sub-series)	Description Level
B2	Title	Mandatory.	
B3	Scope and Content	Optional.	
B4	Reference Code	Mandatory. Must be a unique value, not repeated within your data.	
B5	Year from	Mandatory. Year only.	

B6	Year to	Mandatory. Year only.	
B7	Extent/ Quantity	Optional, but if column is used, the following rules apply: (1) permitted terms are as follows: file, volume, photograph, document [*in reality the list of permitted terms is likely to be longer]* (2) cell must contain no more than 40 characters	
B8	Access restriction	Mandatory. Must be either 'Open' or in the form 'Closed until 2015'	
B9	Notes	Optional.	
X1	[no obvious equivalent]		Accession number
X2	[no obvious equivalent]		Date details
X3	[no obvious equivalent]		Box / location

DATA MAPPING EXERCISE: SUGGESTED ANSWERS FOR TABLE B (Rules for Individual Elements

	ARCHISEARCH Element Name	***ARCHISEARCH Element Rules***	***Equivalent element(s) in data sample (QUINCE workbook)***	***Data mapping rules***	***Using Excel to implement data changes***
B1	Level of Archival Description	Mandatory. The following terms are permitted: Fonds; Series; file; item (together with 'sub-levels such as sub-series)	Description Level	(1) change column heading to 'Level of Archival Description' (2) series and item are permitted terms: no change (3) 'Section' is not a permitted term: consider changing to sub-fonds (or seek advice from system owners) (4) mandatory field: check that there are no blank cells in this column *NOTE: sample data only contains one example of a non-permitted term (D7338/11), so it may be worth examining a larger sample to see if there are any more examples which may have been missed.*	(1) filter to view a list of terms and identify anomalies/ blank cells (2) Find and replace to convert 'Section' to permitted term.
B2	Title	Mandatory.	Title	Mandatory field: check that there are no blank cells in this column. (There is only one example: D7338/10/9/5)	Use filter and view only blanks; insert title manually as required.
B3	Scope and Content	Optional.	Description	change column heading to 'Scope and Content'	

	ARCHISEARCH Element Name	***ARCHISEARCH Element Rules***	***Source data equivalent***	***Data mapping rules***	***Using Excel to implement data changes***
B4	Reference Code	Mandatory. Must be a unique value, not repeated within your data.	Reference Number	(1) mandatory field: check that there are no blank cells in this column (2) there should be no repeated reference numbers, but check first before carrying out the migration (Note: there are no repeated reference numbers in the sample)	(1) filter to check for blank cells (2) this formula looks for duplicates in cells A1 to A20: =COUNTIF($A$1:$A$20,$A$1:$A$20) Change A1 AND 20 to reflect the range being checked, and paste the formula into every row (3) alternatively, copy the reference number column, paste it into a new location, use 'remove duplicates' and see if the number of rows reduces.
B5	Year from	Mandatory. Year only.	Covering Years	(1) Covering Years is mostly in the form year from – hyphen – year to. 'Year from' is the value before the hyphen. (2) UNRESOLVED ISSUE: Rule provided does not give enough detail. Need to ask system owners what to do if there is only one date (eg D7338/10/4/8): is 'year from' data repeated in 'year to'? (3) UNRESOLVED ISSUE: Ask if 'circa' dates and other approximations are permitted (eg 'c. 1900'; 'nineteenth century').	Standard dates in the form year from – hyphen – year can be separated using 'Text to Columns', separating the data at the hyphen. However, a careful check and perhaps some manual editing may also be necessary for anomalous data.
B6	Year to	Mandatory. Year only.	Covering Years	Covering Years is mostly in the form year from – hyphen – year to. 'Year from' is the value after the hyphen. Other queries in 'year from' also apply to 'year to'	

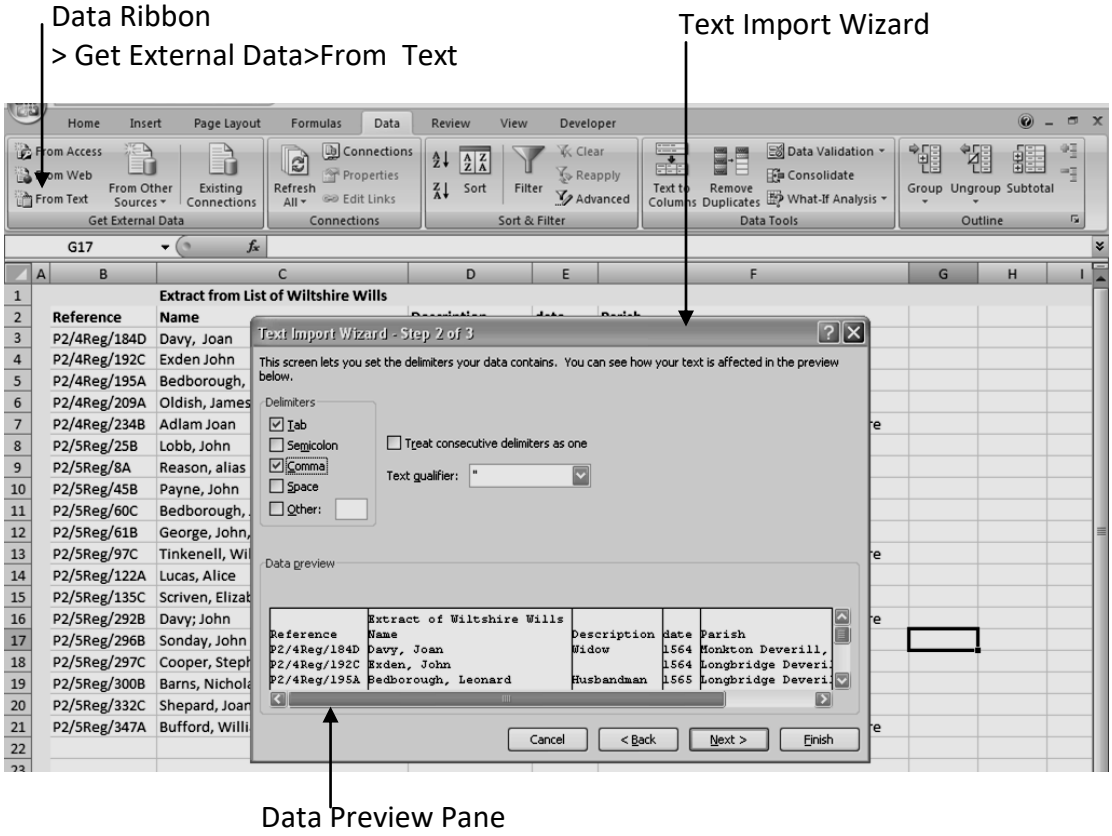
B7	Extent/ Quantity	Optional, but if column is used, the following rules apply: (1) permitted terms are as follows: file, volume, photograph, document (2) cell must contain no more than 40 characters	Extent	(1) Vol is not a permitted term: change to 'Volume'. (2) 'Bundle' and 'rolls' are not on the list of permitted terms: ask system owners for guidance on whether these may be retained. (3) Check that no cells contain more than 40 characters. Note that the sample contains only one instance of this being exceeded D7338/11/1/2 – so it may be worth examining a larger sample to see how many instances there are likely to be [*in reality the list of permitted terms is likely to be longer]*	(1) Find and replace (or the VLOOKUP function if a long list of terms need to be replaced). (2) the LEN function returns the number of characters in a cell.
B8	Access restriction	Mandatory. Must be either 'Open' or in the form 'Closed until 2015'	[Data not held]	This is a mandatory column, so a new column must be inserted and data created. Populate each row with the value 'Open' but before carrying out the migration examine the data to confirm that no Closed records are included.	Copy and paste 'Open'
B9	Notes	Optional	[Data not held]	(1) Optional, so do not need to populate this column (2) alternative: use for Accession number.	
X1	[no obvious equivalent]		Accession number	(1) Data for internal use only; do not include in migration (2) alternative: put in the 'notes' field, preceded by the words 'Accession Number', eg '9445' becomes 'Accession Number 9445'	For option 2, Concatenate function
X2	[no obvious equivalent]		Date details	Add to scope and content column, after the data from the existing 'Description' column. [or could use the notes column]	Concatenate 'Description' and 'Date Details', perhaps with a full stop in between, or the words 'date details'. For example: =CONCATENATE(H4,". Date details ",G4)
X3	[no obvious equivalent]		Box / location	Data for internal use only; do not include in migration	

<< END OF THIS UNIT >>

UNIT N1
IMPORTING INTO EXCEL USING CHARACTER SEPARATED VALUES

Purpose: To move data between Excel and other systems

Excel Workbook: This unit uses the NECTARINE workbook
Text Document: The unit also uses the **Text** document NECTARINE.TXT, which contains a brief extract from a catalogue
Text Editor: the Unit also uses a Text Editor (explained in section B of this Unit)



***NOTE**: This illustration shows Excel 2007; later versions have a slightly different appearance*

A INTRODUCTION AND OVERVIEW

1 **Overview:**

- In this exercise data is imported into Excel from a text file in CSV format, a format frequently used for moving data between databases
- 'CSV' stands for 'character-separated values' (or, formerly, for 'comma-separated values'); it may also be referred to as 'character/comma delimited format'
- Excel acts as the driving force, 'pulling in' the data
- Note that it is not possible to reverse the import of data using the Excel 'undo' function

2 **Learning points:**

- Importing data into Excel as an example of a data migration
- Use of a text editor program to view and edit data
- Introduction to data in CSV format
- Awareness of issues which may arise when migrating special characters (for example , or /)
- Awareness of the degree of control and selection which may be available when importing data

B USING A TEXT EDITOR

3 In this UNIT you need to use a type of program called a Text Editor

- A Text Editor allows editing of text, but little or no formatting, resulting in small file sizes and low risk of any automatic formatting which might interfere with a data migration
- The Text Editor does not have advanced functionality which would permit, for example, importing or exporting data.

4 Most computers have a Windows program called NOTEPAD, which can normally be found at All Programs > Windows Accessories > Notepad.

5 More sophisticated Text editors are available, which can be easier to use than Windows Notepad if you need to carry out a lot of editing.

- They are often available as free downloads, for example *Notepad++* (https://notepad-plus-plus.org/).

6 Text Editors can open files in a variety of formats, but mainly *.txt, *.csv and *.xml

- To open files in formats other than .txt, it may be necessary to choose 'Files of Type' > All Files when browsing to view available files

7 Although it is possible to use a text editor to open files in more complex formats such as Excel or Word documents, the result is meaningless

8 If no text editor is available, it is possible (though not ideal) to use MS Word for the purposes of the exercises in UNIT N1

C MIGRATION EXERCISE 1: CSV format into Excel

9 Using a TEXT EDITOR such as NOTEPAD, open the document entitled **Nectarine.txt**

- In the list of files, right click on Nectarine.txt and choose Open with Notepad
- Alternatively within Notepad or other text editor, select File > Open
- The suffix *.txt indicates a text document

10 In the TEXT EDITOR, view the file **Nectarine.txt**

- This is the data which will be imported into Excel
- The file contains a short extract from a catalogue of Cooperative Society records
- The data may have been created by typing directly into the text editor or may have been generated as an export from another system.
- Although not laid out as a table, some structure is evident, with different data elements separated by semi-colons

11 In EXCEL, open the document 'NECTARINE.xlsx'; this is the blank document into which the data will be migrated
Go to Worksheet 2 (which is empty)

12 On the Data Ribbon, 'Get External Data' Group, select 'From text'
In the dialogue box choose the file Nectarine.txt
At the bottom of the dialogue box Click 'Import'

- Step 1 of the Text Import Wizard opens

13 In Step 1 of the Text Import Wizard, ensure that **'Delimited'** is selected under 'original data type', and that '**row 1'** is selected for 'start import at row'.

- Note that the 'file origin' box only applies to data created using special character sets, such as Cyrillic text.

14 Click 'Next'

- Step 2 of the Text Import Wizard opens
- In Step 2 of the Text Import Wizard, note that a preview of the migrated data appears at the foot of the import Wizard dialogue box.

15 In Step 2 of the Text Import Wizard, ensure that both 'tab' and 'semicolon' are selected

- Note that in the preview, the column structure changes when different options are chosen (eg selecting / unselecting 'semicolon')

Ensure that 'Treat consecutive delimiters as one' is NOT selected
The 'text qualifier' need not be changed
Click 'Next'

- Step 3 of the Text Import Wizard opens

16 In the preview of Step 3 of the Text Import Wizard, select each column in turn.
Ensure that each column is formatted as 'Text' or 'General' data type
Ensure that no column is marked as 'do not import column (skip)'
Click 'Finish'

- Import Data Dialogue Box appears

17 In the Import Data Dialogue Box select 'existing worksheet'

- The cell address shows where the imported data will be located (any cell may be chosen; A1 or $A$1 indicates the top left of the current worksheet)
- Choosing 'new worksheet' will place the data in a new worksheet within the existing workbook

18 Click OK

- The data from Nectarine.txt appears within Excel
- Note that the semi colons which separated the data elements no longer appear; they have been replaced by the column structure

D MIGRATION EXERCISE 2 (optional): More practice migrating CSV format into Excel

19 **Overview:**

- This optional exercise provides additional practice in migrating data from the CSV file into Excel
- It repeats the steps in section C above, but experimenting with additional options, including editing data using the text editor and using different delimiter characters

20 Using a TEXT EDITOR such as NOTEPAD, re-open the document entitled **Nectarine.txt**
Within the text editor, make some changes to the data. For example, change one or more of the years to 2016 (this is to demonstrate that changes made in the text editor will be reflected in the data once migrated to Excel)

- Note that adding or deleting semicolons may alter the data structure

21 In the text editor, save the file as **Nectarine copy 1.txt**

22 In EXCEL, open a new worksheet within the workbook NECTARINE.xlsx'

23 On the Data Ribbon, 'Get External Data' Group, select 'From text'
In the dialogue box choose the file Nectarine copy 1.txt
Click 'Import'

24 In Step 1 of the Text Import Wizard, select **'Delimited'** and 'row 1' as before
- You may experiment with changing these options, but if so, return to using 'delimited' before proceeding to step 25

25 In Step 2 of the Text Import Wizard, choose COMMA as the delimiter
In the preview, observe what happens to the data
- The data is divided differently from using the semicolon and because some names have different numbers of commas (see for example 'Register of Members, Tetbury area, with separate index' in row 5)
- Excel does not distinguish between the comma used within text or as a delimiter; using a comma as a delimiter is often undesirable with textual data

26 OPTIONAL: If you wish, complete the import (choosing 'finish') and observe what the data looks like in Excel; you will then need to restart the import (steps 23 and 24) before proceeding.

27 In Step 2 of the Text Import Wizard, choose OTHER as the delimiter
Type a forward slash (/) in the box next to Other
In the preview, observe what happens to the data
- Excel does not distinguish between the slash used within text (in this case in the reference number) or as a delimiter

28 In Step 3 of the Text Import Wizard, experiment with marking one or more columns as 'do not import column (skip)' before completing the migration by clicking 'finish'

29 In the Import Data Dialogue Box change the location into which the data will be migrated, either by changing the cell address or by selecting 'new worksheet'

30 Within Excel, observe the effect that the data changes and different options have made to the structure of the data. These will depend on the editing carried out and the options chosen, but may include the following:
- Changes made to the data using the text editor are reflected in the migrated data
- Using different delimiters affects the column structure

<< END OF THIS UNIT >>

UNIT N2
EXPORTING FROM EXCEL USING CHARACTER SEPARATED VALUES

Purpose: To move data between Excel and other systems

Excel Workbook: This unit uses the OLIVE workbook
Text Editor: the Unit also uses a Text Editor (explained in section B of Unit N1)

A Exporting from Excel into tab-delimited format

➔ *Open the **OLIVE** Workbook*

1 **Overview:**
- In this exercise data is exported from Excel into a text file in tab-delimited format (the columns in Excel are replaced by the tab character)
- This time, Excel acts as the driving force, 'pushing out' the data

2 **Learning points:**
- Increased understanding of Character Separated Values format
- Exporting data from Excel as an example of a data migration
- Awareness that when exporting there may be less control over the data and format then when importing
- Use of "text qualifiers"

3 In Excel, open the OLIVE Workbook, Worksheet 'Sample 1'; this contains a short extract from a list of Wills for two Wiltshire parishes (a different extract form that used in Unit N1) .

4 Make a few changes to 'personalise' the data in order to demonstrate that the changes are reflected once the data is migrated. For example, change one or two of the dates to the current year, and one or two of the parishes to your own home town.

5 Save the Workbook using File > Save as > Other Formats
In 'Save as Type', the following option from the pull down list:
Text (tab delimited)(*.txt)

6 Click Save
- Message appears explaining that this format cannot be used to save multiple worksheets

7 Click OK; this will save the active sheet only
- Note that the file name at the top of the screen changes to **OLIVE.txt**
- Message appears explaining that the workbook may contain features incompatible with Text (tab delimited) format

8 Click yes
Save and close the workbook
- If the 'workbook may contain features incompatible with Text (tab delimited) format' message is displayed again, choose 'yes' a second time.

9 Open the file OLIVE.txt using either Word or a text editor such as NOTEPAD
Examine the data to confirm that the changes made in step 4 are reflected in the migrated data

- For Notepad and other text editors, see UNIT N1 section B
- Make sure you open OLIVE.txt (not the Excel document OLIVE.xlsx)
- To select the file in Word, it may be necessary to select **Files of type > All files (*.*)**

10 Observe how the column structure has been translated into the text file

- Note that the separate columns of data have been replaced by tab characters
- In Notepad the tab characters appear as spaces
- In Word, the tab characters can be seen by choosing Home ribbon > Paragraph Group > show paragraph marks and other hidden formatting symbols

11 Observe that some data has been marked with quotation marks (“ ”)

- Where a cell contained a comma, “ “ marks are inserted automatically as a ‘text qualifier’ to prevent the commas being treated as delimiters.

12 In order to check the effectiveness of the migration, it is possible to import this new file **OLIVE.txt** back into Excel (open a new Excel worksheet and follow the procedure used in UNIT N1)

- In step 2 of the Text Import Wizard, select tab as the delimiter, and set the text qualifier to double quotes (“).
- When the import is completed quote marks disappear, and the data moves into the correct columns
- If the procedure is repeated setting the text qualifier to ‘none’, the double quotes are treated as text, and are imported.

B Exporting from Excel into comma delimited format

➔ *Open the **OLIVE** Workbook*

13 **Overview:**

- In this exercise data is exported from Excel into CSV, or comma-delimited, format
- Excel acts as the driving force, ‘pushing out’ the data

14 **Learning points:**

- increased understanding of character delimited formats
- Exporting data from Excel as an example of a data migration
- Awareness that when exporting there may be less control over the data and format then when importing
- The importance of testing a migration by re-importing back into Excel

15 In EXCEL, open the document 'OLIVE.xlsx', worksheet 'sample 1'. This contains a short extract from a list of Wills for two Wiltshire parishes.

16 Make some changes to the data to 'personalise' it, in order to demonstrate that the changes are reflected once the data is migrated

17 Save the Workbook using File > Save as > Other Formats
In 'Save as Type', select the following option from the pull down list:
CSV (comma-delimited)(*.csv)

18 Click Save
- Message appears explaining that this format cannot be used to save multiple worksheets

19 Click OK; this will save the active sheet only
- Note that the file name at the top of the screen changes to **OLIVE.csv**
- Message appears explaining that the workbook may contain features incompatible with CSV (comma delimited) format

20 Click yes
Close the workbook
- If the 'workbook may contain features incompatible with CSV (comma delimited) format' message is displayed again, choose 'yes' a second time.

21 Open the file OLIVE.txt using either Word or a text editor such as NOTEPAD
Examine the data to confirm that the changes made in step 16 are reflected in the migrated data
- For Notepad and other text editors, see UNIT N1 section B
- Make sure you open OLIVE.csv (not the Excel document OLIVE.xlsx)
- To select the file in either Word or Notepad it may be necessary to select **Files of type > All files (*.*)**

22 Observe how the column structure has been translated into the text file
- Note that the separate columns of data have been replaced by commas

23 Observe that some data has been marked with quotation marks (" ")
- Where a cell contained a comma, " " marks are inserted automatically as a 'text qualifier' to prevent these commas being treated as delimiters.
- The automatic process of distinguishing between commas used as text and as delimiters may not always be perfect and is not easy to check in the text document; it is easier to check the migration by importing back into Excel

24 In order to check the effectiveness of the migration, it is possible to import this new file OLIVE.csv back into Excel (using the procedure in UNIT N1)
- In step 2 of the Text Import Wizard, select comma as the delimiter, and set the text qualifier to double quotes (").
- When the import is completed quote marks disappear, and the data moves into columns.

C Character separated formats hints and tips

25 Importing an exported file back into the original system is a good way to test the effectiveness of the process.

26 When exporting to a *.CSV format, the separator character is always a comma. It is only possible to change this by changing the default character for all programs on the computer. This may have adverse consequences for other programs, and you may not have the permissions to make such changes.

27 Comma delimited format can be required as a means of transferring data into some systems, but when applying to text which includes commas it is not always obvious if the text qualifier (double quotes surrounding fields including commas) have been applied correctly. If there are problems with CSV, the following techniques may work (test them carefully; they may work differently in different circumstances).

- Save as tab delimited format, which may be more effective at inserting 'text qualifiers'. In Word or a text editor, use find and replace to change all the tabs to commas.
- Within Excel, use find and replace to change all the commas to an unused character (such as #). Save as a CSV (comma-delimited)(*.csv) document. Migrate the data from *.csv into the final system, then find and replace all the # back to commas. NOTE: this depends on the final system supporting search and replace.

MORE PRACTICE

EXERCISE 1

Open the OLIVE workbook, Worksheet 'Sample 2'. This contains a short extract from the catalogue of a business archive. Experiment with exporting from Excel using tab delimited format and comma delimited format. Observe how the data appears in each format, and particularly how the migration affects the numerous punctuation marks.

EXERCISE 2

Using the .txt and .csv files created in Exercise 1, experiment with importing the data back into Excel.

<< END OF THIS UNIT >>

UNIT N3
INTRODUCING XML and EAD

This UNIT mainly covers the theory, but includes some (optional) practical exercises

A What is XML?

1 This section is aimed at helping you to understand documents written in XML. It does not cover creating XML documents from scratch. For EAD (Encoded Archival Description), which is a specific version of XML, with additional rules, see Section B

2 The acronym XML stands for Extensible Markup Language:
Markup because existing text (eg a catalogue entry) is *marked up* by adding XML codes or 'tags'.
Language because XML has a specific syntax and structure which must be followed
Extensible because you can extend it by creating your own tags and adding as many as you like (within the rules)

3 XML is related to HTML (Hypertext Markup Language) used to build websites, and has a similar structure, but the two are not (normally) interchangeable.

4 Example of a fragment of XML:

<reference>D2754/8/16**</reference>**
<title>Cooperative Women's Guild Poster. Women's Cooperative Guild Central Committee.**</title>**
<scope and content>Shows the committee's organisational structure**</scope and content>**
<date>1949**</date>**

5 The descriptive phrases between the angle brackets (such as <title>) are known as 'tags', shown in bold above. They work in pairs: an opening tag at the start of a particular data element, and a closing tag at the end. The opening and closing tags are identical except for a slash to indicate the closing tag (compare <date> and </date>).

6 The purpose of XML is to classify different types of data element using text alone, independently of any layout structure or software. It can therefore be used to transfer data between systems. In the fragment above there is sufficient information

- To create a Word table, or an Excel document.
- To import the data into a database structured with four fields (reference, title, scope and content, date).
- To distinguish between title and scope-and-content data: Without any tags, you might think "*Cooperative Women's Guild Poster. Women's Cooperative Guild Central Committee. Shows the committee's organisational structure"* is all the title, or that the title ends with the word 'Poster'.
- Using appropriate software, to display or search for only the one type of element (eg title), where multiple records are tagged in the same way.

7 Because XML follows specific rules for structure and syntax (not covered in any detail here), it can be used and interpreted by a huge range of software, including Excel, Web browsers and database import tools. There are also many tools available for creating an XML document: it is not normally necessary to write the XML code by hand.

8 Example of a complete XML document containing two records

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<data-set>

<record>
<reference>D2754/8/16</reference>
<title>Cooperative Women's Guild Poster: Women's Cooperative Guild Central
Committee</title>
<scopeandcontent>Shows the committee's organisational structure</scopeandcontent>
<date datetype="yearonly">1949</date>
</record>

<record>
<reference>D2754/8/17</reference>
<title>Cooperative Women's Guild Poster: Join the Cooperative Women's Guild</title>
<scopeandcontent> </scopeandcontent>
<date datetype="yearonly">1960</date>
</record>

</data-set>
```

9 Features of the XML document illustrated:

- The first line is called the XML declaration. This is essential to make clear to any software processing the data that this is an XML document
- The first tag <data-set> defines the 'root element', and is balanced by the final tag </data-set>. An XML document must have a root element. All the other data is nested within the root element tags.
- Other tag pairs may also be nested (for example <record> </record> has the reference, title, scopeandcontent and date elements nested within it)
- Tags must be lowercase only
- The spaces between the data for each record are not essential, and are only for visual clarity. Neither is it essential to start each tag on a new line.
- Tag pairs need not enclose data (see for example the second <scopeandcontent></scopeandcontent) *[this can also be coded with an 'empty tag' <scopeandcontent /]*
- The <date> tag contains some additional information: **datetype="yearonly**", which can vary (for example datetype="monthandyear"). This is called an 'attribute'. The attribute need not be repeated in the closing tag </date>
- Although this is a valid XML document and contains archival descriptions, it does NOT meet the requirements of EAD (Encoded Archival Description), for which see section B of this UNIT

B What is EAD?

10 EAD is an international standard developed by the Society of American Archivists and the US Library of Congress. It maps closely to the International Standard for Archival Description ISAD(G)

11 EAD uses XML format and a specific set of tags and syntax

12 The relationships between levels of description are expressed by using tags indicating 'components' which are nested within each other

13 Although EAD is a standard, it does come in different versions which differ slightly, so it is important to know which version is required in a particular context

14 More information about EAD as used by the Archives Hub is available on the Hub website:
http://archiveshub.ac.uk/ead/

15 More information on EAD generally, including a complete list of permitted tags, is available from the EAD official website:
https://www.loc.gov/ead/

C Creating XML and EAD files

16 Although it is possible to create valid XML by typing the tags manually, this can be time consuming and prone to error.

17 Special software is available for creating XML, which generates the tags for you, and the Archives Hub (for example) provides a specialist EAD editor.
See for example:
http://archiveshub.ac.uk/eadeditor/
http://archiveshub.ac.uk/xmlsoftware/

18 More detail on creating EAD files is beyond the scope of this workshop.

D Viewing XML and EAD files

- This practical section is **Optional**
- This section uses the XML file PINEAPPLE.XML

19 XML (and EAD) files can be viewed in a variety of software, which demonstrate different features, including:

- Text Editors
- Web Browsers
- Word
- Excel

20 Although the detail of viewing and editing XML files is beyond the scope of this Workshop, the following steps provide a taste of what is involved

- You may OMIT any or all of the following exercises

21 **Exercise 1 viewing XML in a Text editor**
Using a Text Editor such as Notepad open and view the file PINEAPPLE.xml (for text editors see UNIT N1, section B)

- This contains an extract from a catalogue with the XML tags displayed
- It is similar in format and structure to the example shown in step 9 above

22 **Exercise 2 viewing XML in a Web Browser**
Using a Web Browser (such as Internet Explorer, Firefox, or any other), open and view the file PINEAPPLE.xml

- This contains an extract from a catalogue with the XML tags displayed
- It is similar in format and structure to the example shown in step 9 above
- What you see may differ depending on the web browser used and its settings (for example, the declaration at the start of the XML document may or may not display)

23 **Exercise 3 viewing XML in Word**

Using Word, open and view the file PINEAPPLE.xml

- This contains an extract from a catalogue
- The XML tags and the declaration at the start of the XML document may or may not display (depending on the version and settings of Word)

24 **Exercise 4 viewing XML in Excel**

Open Excel

In Excel, open and view the file PINEAPPLE.xml

In the 'Open XML' dialogue box choose 'As an XML Table' and click OK

If a message 'The specified xml source ....' appears, click OK to close it

- This contains an extract from a catalogue
- The XML tags do not display
- The document has been transformed into a (fairly) normal-looking Excel document

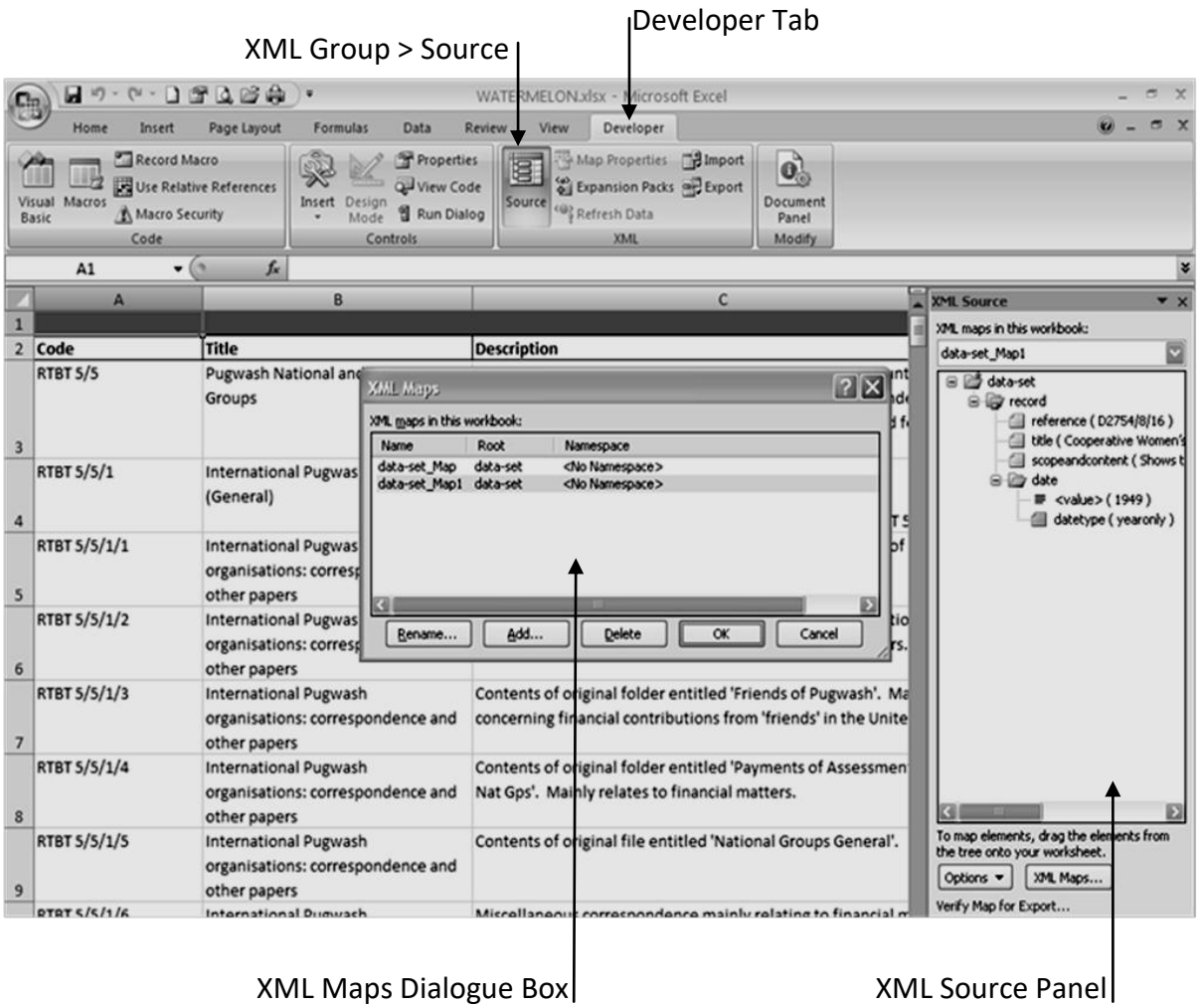
25 Note that the Excel document is NOT called 'PINEAPPLE' : opening the document in Excel has resulted in the creation of a NEW Excel document

26 The new document can be saved (if you wish) as a normal Excel document using 'save as'

<< END OF THIS UNIT >>

UNIT N4
EXPORTING FROM EXCEL TO XML

Excel Workbook: This unit uses the WATERMELON workbook
XML document: it also uses the document WHITECURRANT.XML



A FIRST STEP: ENABLE THE DEVELOPER TAB

1 Open Excel
Look at the Excel Tabs near the top of the screen (such as 'Home', 'Formulas' 'Data')
If there is one called 'Developer', the tab has already been enabled, so omit the rest of section A and go to section B

2 The Developer tab is not normally displayed by default. If it is not visible it must be enabled.

3 The Developer tab is selected within Excel Options, but the exact steps will depend on the version of Excel you are using.

In Excel 2007:
Click the Office Button (top left of screen) and choose Excel Options
Choose 'Popular' options
In the dialogue box, tick the box 'Show the Developer Tab in the Ribbon'
At the bottom of the Dialogue Box, click OK

In Excel 2010 and later:
Choose File > Options
In the Excel Options dialogue box, choose Customise Ribbon
Tick the 'Developer' box
Click OK

B Overview

4 In this UNIT, data is exported from the WATERMELON Workbook into XML format

5 This requires the following steps:

a) Identify (or create) an XML template
b) Structure the data in Excel columns to match the template
c) Add the template to the Excel document as an XML map
d) In Excel, map the data columns to the template
e) Save the Excel document as an XML document

6 Note that although this method produces a valid XML document and contains archival descriptions, it does NOT meet the more sophisticated requirements of EAD (Encoded Archival Description), for which see UNIT N3

C Identifying an XML template

7 XML documents have the file extension **.xml**
Locate and open the file WHITECURRANT.xml
- By default, the file will probably open in a browser such as Internet explorer
- This file will be used as a template
- Note that this template is the sample XML file illustrated in UNIT N3, step 8

8 Locate the file WHITECURRANT.xml and choose Open With [right click on the file name]
- Note that XML documents can be opened in a variety of programs, including Word, WordPad and other text editors, and in Excel (see UNIT N3)

9 Locate the file WHITECURRANT.xml and choose Open With Excel
- You may get a message "WHITECURRANT.xml is locked for editing ..." if so, choose to open it Read Only
- Depending on how your computer is set up, you may need to open a blank Excel document before choosing 'Open with Excel'
- A dialogue box opens "Please select how you would like to open this file:"

10 Choose the option "as an XML table" and click OK
- You may get a message "The specified XML source does not refer to a schema... ". This can be ignored: Click OK
- The document opens in Excel, with five columns, a header row and two rows of data.
- This is a new and un-named Excel document (it may be renamed and saved as an .xlsx document if you wish, but this is not required for the purposes of this exercise)

Leave this excel document open, so that you can refer to it in the next section

11 Creating a template from scratch is not in scope for the Workshop, and is rarely necessary, since the reason for using XML is normally to move data into a pre-existing format.

D Structuring the data in Excel columns to match the template

12 Open the document WATERMELON.xlsx, Worksheet Copy 1
- This represents part of a catalogue very different in type from that in the XML template (ie WHITECURRANT.xml), which we wish to export into the same XML structure

13 Compare the column headings WATERMELON.xlsx with those in the first row of the template WHITECURRANT.xml, as opened in Excel

- Note that the names of the column headings are not identical, but they are equivalents (for example 'reference' and 'code' both contain the document identification number)
- The equivalent columns are in the same order in both documents; this is not essential, but will make the mapping easier
- Note that the dates are in a completely different format from those in the template (from-to dates instead of a single year) and that different 'date type' terms are used.
- Note that there is an additional column, 'Extent' in WATERMELON.xlsx which has no equivalent in the template and cannot therefore be mapped or exported.

14 In this example, the data has already been structured as required, so it is not necessary to move or change the column order or structure.

FOR INFORMATION:
If the data were to need substantial re-arrangement, a helpful approach could be to use the WHITECURRANT template (as opened in Excel) as the basis for a new document. The new data for export (in this case the WATERMELON data) can be copied column by column, and pasted into the correct column in the new (whitecurrant) document (don't forget to re-name and save it).

E Adding the template to the Excel document as an XML map

15 Open the source document WATERMELON.xlsx
Go to the **Developer Tab**
If you cannot see the Developer Tab, see section A of this UNIT

16 In the **Developer Tab > XML Group**, click 'Source'

- A panel headed 'XML source' opens on the right of the screen

17 In the XML Source Panel, choose 'XML Maps'

- XML Maps dialogue box opens

18 In the XML maps dialogue Box choose 'Add'

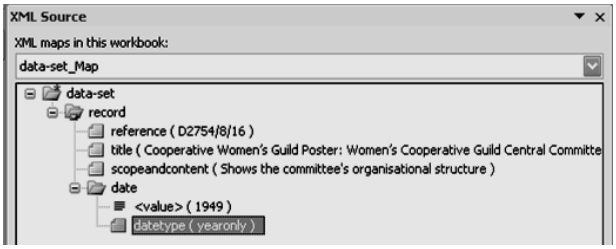
- Select XML source browser box opens

19 Navigate to the file WHITECURRANT.xml
Select it and click 'open'

- You may get a message "The specific XML source does not refer to a schema... ". This can be ignored: Click OK to close it
- an XML map with the root tag 'Data-set' appears in the XML Maps dialogue box

20 In the XML maps Dialogue Box choose OK to close it

- In the XML Source Panel, in the WATERMELON Workbook the column heading names from the template are displayed, showing their relationships in a tree structure
- Depending on the settings, the data from the first record in the template may also display, as illustrated below (if the data is not displayed, choose 'Options' in the XML Source Panel and select 'preview data in task pane')



F	**Mapping the data columns to the template**

21 In the XML Map in the XML Source Panel, click 'scopeandcontent'

- scopeandcontent is highlighted to show that it has been selected

22 Click 'scopeandcontent' and drag to the cell C2 ('Description') on the Excel Worksheet.

- The whole column 'description' is formatted and a filter arrow appears
- Note that although the 'description' column has been mapped to 'scopeandcontent', the column name has not changed (though the text may be reformatted to a pale colour)

23 In the XML Map in the XML Source Panel, click 'reference' and drag to cell A1 (ie the cell above the column heading 'code'

- The whole column A is formatted and a filter arrow appears
- Note that because there was no column heading in cell A1, one has been created, entitled 'reference'. The word 'code' in row B is treated as the first item, of data.
- It is possible to map each column individually in this way, which may be useful if the columns are in an unexpected order, or if some columns (or some elements in the map) are not being used, but it is also possible to make an error this way.
- If you wish, compare the appearance of the mapped data with the unmapped data in WATERMELON.xlsx Worksheet Copy 2

24 Use 'Undo' to undo all the mappings made so far (use 'undo' until all the formatting disappears)
In the XML Map in the XML Source Panel, select the root element ('data-set')
- The root element and all the elements nested below it are highlighted to show that they have all been selected

25 With the root element highlighted, drag it to cell A2 ('code')
- Columns A to E are all formatted, and have filter arrows
- Mapping all the columns together is the safest way to make an accurate mapping, BUT this relies on all the columns being in the correct order
- Note that the 'extent' column has not been mapped as there is no equivalent, [but in practice it would be safer to delete it first]

G Saving the Excel document as an XML document

26 By choosing 'save as' save the document WATERMELON as a new document "WMCOPY.xml", as follows
In the 'save as' dialogue box, choose **save as type: XML Data (*.xml)**
Ensure that the name for the new file is WMCOPY.xml
Click 'Save'
You may get a message "Saving the file as XML data will result in the loss of worksheet features ....". This can be ignored: choose Continue to save the file

27 In your document library (or Windows Explorer) locate the new document WMCOPY.xml
Choose Open With [right click on the file name]
Choose open with Internet Explorer
(alternatives: open with any browser program, or with a text editor such as Notepad or Notepad++)
- The XML tags (and attributes for the date element) have been applied to all of the new data from the document XML EXPORT.xml
- Note that the root element <data-set> has an attribute defining the version of XML, added automatically <data-set xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">

H Error checking

28 The XML Source Panel contains a link 'verify map for export': this can be used to check that the mapping is valid before saving the mapped document as an XML document. If the command reports an error, try re-mapping each field one at a time, using 'verify map for export' each time to discover where the error lies.

29 Opening an xml document in a browser can help identify errors, although browsers do this in different ways. As well as refusing to open a faulty xml document, Firefox, for example, will identify at which point the error lies.

<< END OF THIS UNIT >>

UNIT P1
MIGRATION BETWEEN EXCEL AND SPECIALIST SYSTEMS

INTRODUCTORY

Every system used for cataloguing archives or making those catalogues available has its own method of and requirements for importing and exporting data.

There is no substitute for reading the manuals appropriate to your system, taking part in training provided by the system supplier, and seeking appropriate technical support. The system-specific notes in the following UNITS are NOT intended to provide detailed instructions, merely to provide an overview of what is involved.

If you do not have a separate test system, ask your supplier; this may be included in the price you have already paid for your system.

Remember that only the simplest migrations can include everything; there may well be certain data (for example, some index fields or links between levels of description) which must be added or corrected after migration.

Disclaimer

Although every effort has been made to provide accurate information, data migration is a complex and specialist topic. Different versions of the same system may vary. The notes may not therefore be accurate for your own specific context. They are for general information only and should NOT be relied on as the only or main source of information when carrying out a data migration.

Also please note that these units were last updated in 2018 and some are no longer current.

Systems covered:

UNIT P2: Microsoft Access
UNIT P3: Axiell Calm
UNIT P4: Axiell Adlib
UNIT P5: The archives hub

<< END OF THIS UNIT >>

UNIT P2

Migration between Excel and specialist systems:
MICROSOFT ACCESS

(The following notes apply to Access 2007; other versions of Access may differ slightly)

Excel- compatible transfer formats

- Access can import direct from Excel (as well as other formats)

Location of Import facility

- Access > External Data Tab > Import > Excel

Documentation

- Access Help

Overview of import procedure

- In Access, open the database, but not the table to which the data is to be imported
- Open the External Data Tab ***[see screenshot 1]***
- Use 'browse' to select the Excel file to be imported
- Select 'Append a copy of the records to the Table' and select the Access table to which the data is to be added.
- Click OK to start the import Wizard, which includes a preview of the records ***[see screenshot 2]***

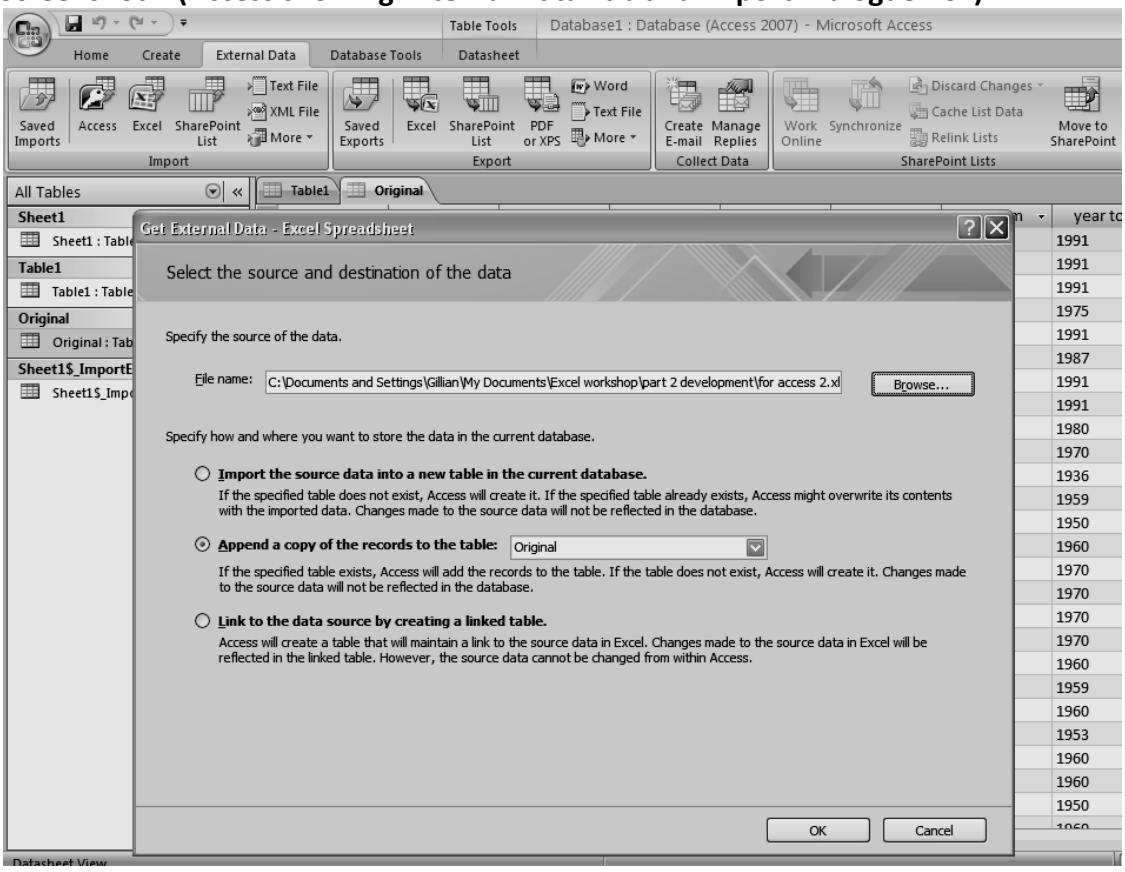
Excel format and structure requirements

- Defining the archival hierarchy: this will depend on the individual Access database, but a parent-child structure is likely, defined by a unique identifier and a parent reference
- The column order and column headings must match exactly those in the destination Access table
- One way of ensuring an exact match is first to export a few records from Access into a new Excel document, and use the column headings in this new Excel document as a template for the source data
- Ensure that there is data in all the mandatory fields, and that data is not duplicated in fields which require unique values (such as a unique identifier)
- For sample data ready for importing into Access, see **Screenshot 3**

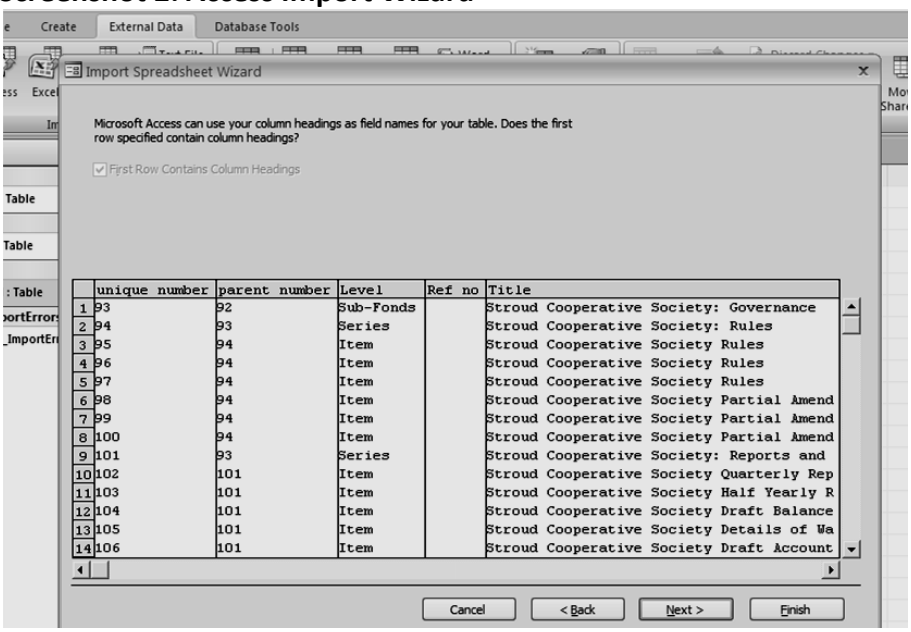
Export from Access to Excel

- In Access open the database and the table which is to be exported
- If only certain records are to be migrated, select them
- Open the External Data Tab
- Choose Export > Excel
- In the dialogue box, choose the file name and location ***[see screenshot 4]***
- Tick 'Export data with formatting and layout'
- Tick 'Open the destination file after the export operation is complete' if required
- Tick 'Export only the selected records' if appropriate

Screenshot 1 (Access showing External Data Tab and Import Dialogue Box)



Screenshot 2: Access Import Wizard



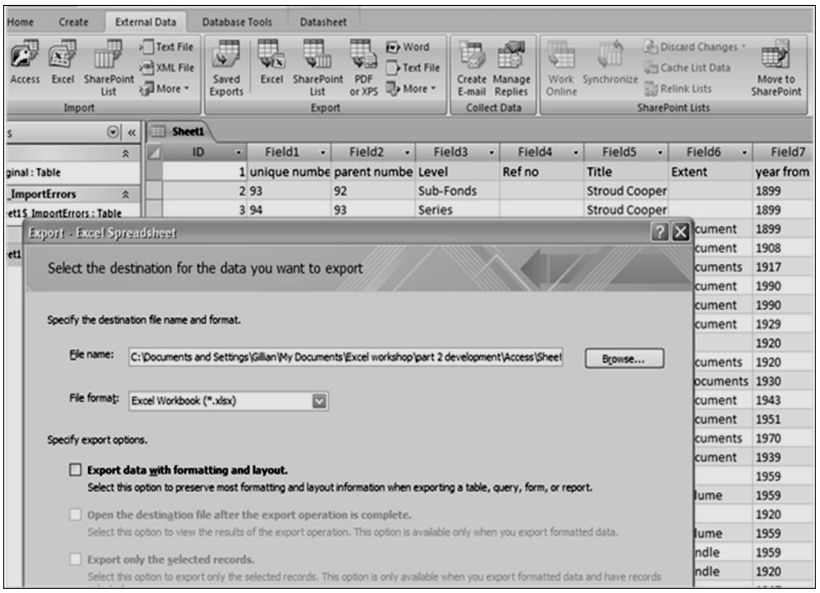
	unique number	parent number	Level	Ref no	Title
1	93	92	Sub-Fonds		Stroud Cooperative Society: Governance
2	94	93	Series		Stroud Cooperative Society: Rules
3	95	94	Item		Stroud Cooperative Society Rules
4	96	94	Item		Stroud Cooperative Society Rules
5	97	94	Item		Stroud Cooperative Society Rules
6	98	94	Item		Stroud Cooperative Society Partial Amend
7	99	94	Item		Stroud Cooperative Society Partial Amend
8	100	94	Item		Stroud Cooperative Society Partial Amend
9	101	93	Series		Stroud Cooperative Society: Reports and
10	102	101	Item		Stroud Cooperative Society Quarterly Rep
11	103	101	Item		Stroud Cooperative Society Half Yearly R
12	104	101	Item		Stroud Cooperative Society Draft Balance
13	105	101	Item		Stroud Cooperative Society Details of Wa
14	106	101	Item		Stroud Cooperative Society Draft Account

Screenshot 3: Data in Excel ready for import into Access

A1

	A	B	C	D	E	F	G	H
1						**Sample data: Access**		
2								
3	***Access field name>***	***unique number***	***parent number***	**Level**	**Ref no**	**Title**	**Extent**	**year from**
4	***Unique number matches next***	**32**	**31**	**Sub-Fonds**	**D275A**	**Stroud Cooperative Society: Governance**		**1899**
5		**33**	**32**	**Series**	**D275A/1**	**Stroud Cooperative Society: Rules**		**1899**
6		**34**	33	Item	D275A/1/1	Stroud Cooperative Society Rules	1 document	1899
7		**35**	33	Item	D275A/1/2	Stroud Cooperative Society Rules	1 document	1908
8		**36**	33	Item	D275A/1/3	Stroud Cooperative Society Rules	2 documents	1917
9		**37**	33	Item	D275A/1/4	Stroud Cooperative Society Partial Amendment of Rules	1 document	1990
10		**38**	33	Item	D275A/1/5	Stroud Cooperative Society Partial Amendment of Rules	1 document	1990
11		**39**	33	Item	D275A/1/6	Stroud Cooperative Society Partial Amendment of Rules	1 document	1929

Screenshot 4: Access Data Export Dialogue Box



<< END OF THIS UNIT >>

UNIT P3

Migration between Excel and specialist systems: **AXIEL CALM**

Excel- compatible transfer formats

- DScribe Natural format
- Comma-delimited (*.csv) or tab-delimited (*.txt)

XML files, including EAD data, can also be imported into Calm, but there is no particular advantage in using this format when moving data from Excel. Importing XML files is not covered below, but is explained in the Calm on-line documentation.

EXPORTING FROM CALM (Overview)

Export formats: DScribe Natural format; Tab or Comma separated format; XML (including EAD).

Accessed from: Calm Catalogue > File > Import

Documentation: online Calm Manual:
http://www.dswebhosting.info/alm/main_menu/importing_and_exporting/exporting_records.htm

IMPORTING INTO CALM

Location of Import facility

- Calm Catalogue > File > Import

Documentation
See the online Calm manual at:
http://www.dswebhosting.info/alm/index.htm

The section on importing starts at:
http://www.dswebhosting.info/alm/main_menu/importing_and_exporting/importing_records.htm

For step by step instructions for importing in DScribe natural format, via a mailmerge into Word, see
http://www.dswebhosting.info/Documents/User%20Guides/Calm%20ALM%20&%20RM%20v10%20Importing%20from%20Spreadsheets.pdf

For supplementary notes on this method, and practice, see page 4 of unit P3, below

Overview of import procedure

- Open the Import records from File dialogue box ***[see screenshot 1]***
- Choose the import method (DScribe Natural or Tab/Comma separated
- Using the 'file' button, navigate to and select the source file (eg a *.csv file)
- Choose the Record Type (eg component)

Excel format and structure requirements

Defining the archival hierarchy: parent-child structure, defined by the Reference Number (RefNo field); for example D275/1 is a child of D275 and parent of S275/1/2

- The names and order of the files is defined by the Record Type being imported. This will normally be Component (for records below fonds level) or Collection.
- When using DScribe Natural format, the mapping of fields in the Excel document to the Calm fields is carried out in Word during the mailmerge process, although this will be easier if the Excel columns are already in the correct order
- When using character separated format, the columns in Excel MUST be in the same order as the fields in your Calm database (which may not be the same as the order of the forms you use for input).
- Before converting to *.CSV format, or tab separated format, remove the column headings (field names) so that they are not imported as data (or convert to *.CSV and then remove the unwanted data in a text editor)

For sample data in Excel see **screenshot 3.**

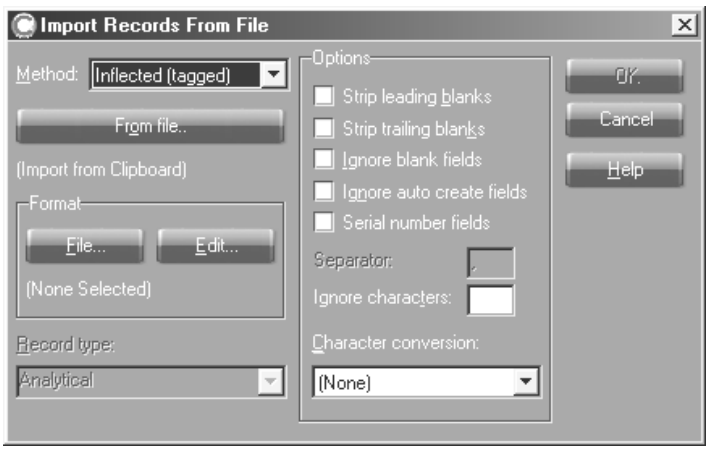
In *.CSV format, the first part of this document in screen shot 3 looks like this:

```
Sub-Fonds,D275A,Stroud Cooperative Society: Governance,,1899,1991
Series,D275A/1,Stroud Cooperative Society: Rules,,1899,1991
Item,D275A/1/1,Stroud Cooperative Society Rules,1 document,1899,1975
Item,D275A/1/2,Stroud Cooperative Society Rules,1 document,1908,1991
Item,D275A/1/3,Stroud Cooperative Society Rules,2 documents,1917,1987
```

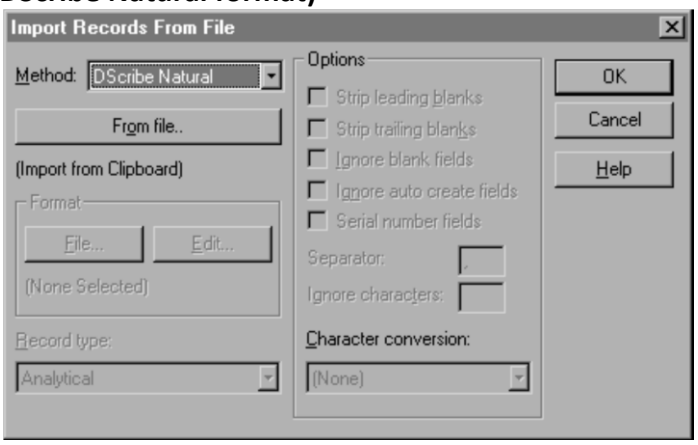
InDScribe Natural format, the first item level record above looks like this:

```
:Component
-Level
Item
-RefNo
D275A/1/1
-Title
Stroud Cooperative Society Rules
-Extent
1 document
-Date Earliest
1899
-Date Latest
1975
```

Screenshot 1 (Import records from file dialogue box)



Screenshot 1 (Import records from file dialogue box showing fields greyed out when using DScribe Natural format)



Screenshot 3: Data in Excel

	A	B	C	D	E	F	G
1				**Sample data: CALM**			
2							
3	*Calm field >*	*Level*	*RefNo*	*Title*	*Extent*	*Date Earliest*	*Date Latest*
4							
5	*Fonds level record NOT included in Component type>*	**Fonds**	**D275**	**Stroud Cooperative Society**		**1899**	**1991**
6	*First record in component >*	**Sub-Fonds**	**D275A**	**Stroud Cooperative Society: Governance**		**1899**	**1991**
7		**Series**	**D275A/1**	**Stroud Cooperative Society: Rules**		**1899**	**1991**
8		Item	D275A/1/1	Stroud Cooperative Society Rules	1 document	1899	1975
9		Item	D275A/1/2	Stroud Cooperative Society Rules	1 document	1908	1991
10		Item	D275A/1/3	Stroud Cooperative Society Rules	2 documents	1917	1987
11		Item	D275A/1/4	Stroud Cooperative Society Partial Amendment of Rules	1 document	1990	1991
12		Item	D275A/1/5	Stroud Cooperative Society Partial Amendment of Rules	1 document	1990	1991
13		Item	D275A/1/6	Stroud Cooperative Society Partial Amendment of Rules	1 document	1929	1980
14		**Series**	**D275A/2**	**Stroud Cooperative Society: Reports and Balance Sheets**		**1920**	**1970**
15		Item	D275A/2/1	Stroud Cooperative Society Quarterly Report and Balance Sheet	3 documents	1920	1936
		Item	D275A/2/2	Stroud Cooperative Society Half Yearly Reports	12 documents	1930	1959

CALM MAILMERGE PRACTICE

This page supplements the notes provided by Axiell entitled Calm v10 Importing from Spreadsheets (http://www.dswebhosting.info/Documents/User%20Guides/Calm%20ALM%20&%20RM%20v10%20Importing%20from%20Spreadsheets.pdf). It also provides a template which can be used for practice when CALM itself is not available. The template simulates the results of the first stage of the process for importing data into Calm from Excel, as documented in the section entitled 'To Create a Template to Map Your Data into' (Axiell notes page 5). Use the notes starting at 'To Add the Spreadsheet Data to the Template' (Axiell notes page 8).

The following Excel for Archivists workbooks are suitable for merging with the sample template:

- MAROON.docx worksheet sheet 'original data' (used in Level One)
- WATERMELON.xlsx worksheet 'sheet 1' (used in Level Two).

Mailmerge is a feature of Word, not of Excel, so is not strictly in scope for the Excel for Archivists Workshops; the following notes provide outline guidance only to supplement that supplied by Axiell:

- Before starting, copy the template below (ie everything between the arrows). Type exactly as shown. Save the word document (for example as 'Calm template 1.docx).
- Take a copy of the Excel workbook from which you will be taking the data. If possible, ensure that row 1 contains the headers (if using copies of MAROON.docx or WATERMELON.xlsx delete row one)
- Once you start the mailmerge process, the Word template and the Excel Workbook become linked together, and can no longer be edited independently, so ONLY use the copies which you have made for the process
- Open the chosen Excel workbook before you start the process, and don't close it or the Word document until the mailmerge is complete.
- If you make a mistake, it is probably best to take a new copy of both the template and the Excel Workbook and start the whole process again.
- If the Excel Workbook contains more than one worksheet, then you will need to specify the worksheet after choosing the Excel workbook (at the point in the Calm instructions reading 'Navigate to the spreadsheet with the data you wish to import')

TEMPLATE

Copy the text between the arrows into a blank Word Document, including : and - symbols
This will serve as the template (it simulates the template generated by exporting from CALM)

↘

:Component
-IDENTITY

-Level

-RefNo

-Extent

-Title

-Date

-Content

↗

<< END OF THIS UNIT >>

UNIT P4

Migration between Excel and specialist systems: **AXIEL ADLIB**

Excel- compatible transfer formats

- Comma-delimited (*.csv)

XML files can also be imported into Adlib, but there is no particular advantage in using this format when moving data from Excel. Importing XML files is not covered below, but is fully documented in the Adlib documentation.

Location of Import facility

- Adlib Designer > Import Jobs Manager

Beware: *Adlib Designer is a tool box for IT professionals and can be used to make irreversible changes to the data, including deleting your entire database.*

Documentation

- Adlib Designer Help 7.2.pdf (starting on page 510)

http://www.adlibsoft.com/support/manuals/maintenance-guides/designer-help-72

Overview of import procedure

- Open Adlib Designer > Import Jobs Manager
- Create, name and describe a new Import Job (don't forget to save it)
- Using the Import Jobs Editor, select the input file type [ASCII delimited (*.csv)]
- Enter details of the source file (including its location path) and the database receiving the data – normally 'catalogue' ***[screenshot 1]***
- Map the Adlib fields the data is to be imported to ***[screenshot 2]***
- Save the import job
- Run the import job

Note that once the import routine has been saved, it can be used again – but ONLY if the Excel structure remains unchanged.

Excel format and structure requirements

- Defining the archival hierarchy: parent-child structure. The Reference Number (eg D275A/1/5) serves as the unique identifier for each record. The 'part of' field contains the reference number of the immediate parent.
- It is also possible to import records as a flat file without a hierarchical structure and build the links within Adlib

- Before converting to *.CSV format, remove the column headings (field names) so that they are not imported as data (or convert to *.CSV and then remove the unwanted data in a text editor)
- When mapping the data to the Adlib fields, Adlib uses column NUMBERS not the field names or Excel column letters. For example, column A is column 01.

See sample data in the Excel file (screenshot 3). The field mapping for this file would be as follows:

Source Field	*Destination Field*	*[Comments: optional]*
01	bt	Parent
02	gv	Level of description
03	IN	Ref number
04	TI	Title
05	DA	Extent
06	DS	Year from
07	DE	Year to

In *.CSV format, the first part of this document looks like this:

```
,Fonds,D275,Stroud Cooperative Society,,1899,1991
D275,Sub-Fonds,D275A,Stroud Cooperative Society: Governance,,1899,1991
D275A,Series,D275A/1,Stroud Cooperative Society: Rules,,1899,1991
D275A/1,Item,D275A/1/1,Stroud Cooperative Society Rules,1 document,1899,1975
D275A/1,Item,D275A/1/2,Stroud Cooperative Society Rules,1 document,1908,1991
D275A/1,Item,D275A/1/3,Stroud Cooperative Society Rules,2 documents,1917,1987
```

Export from Adlib

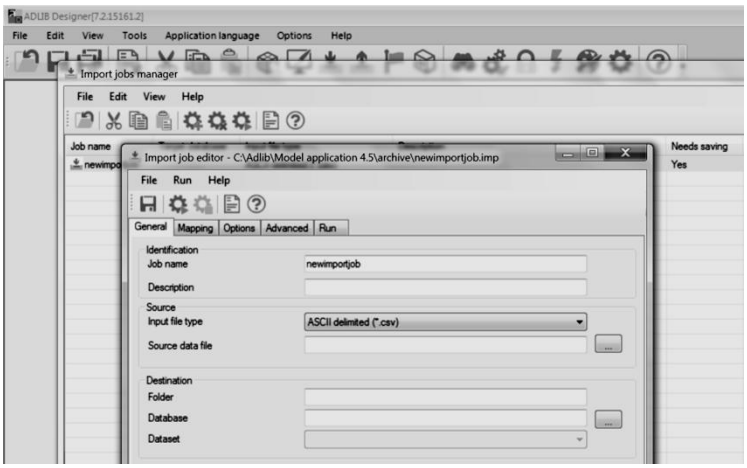
Export formats: include *. csv and XML.

Accessed from: The procedure is similar to importing: an Export job is created, defined and then run using Adlib Designer > Export Job Manager.

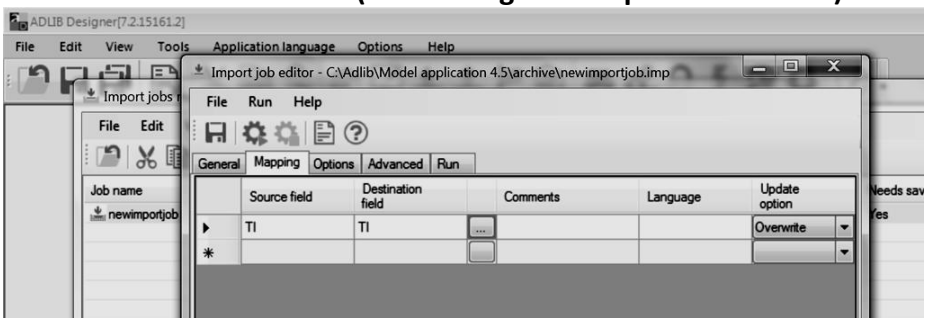
BEWARE: If a field is repeated (for example, if 'Extent' has been used more than once for a single record) only the first instance is exported when using *.CSV format. It may therefore be preferable to use XML format if fields are repeated

Documentation: Adlib Designer Help 7.2.pdf (starting on page 510)

Screenshot 1 (Adlib Designer > Import Jobs Editor)



Screenshot 2 Screenshot 1 (Adlib Designer > Import Jobs Editor)



Screen shot 3 (Excel before saving as *.CSV)

	A	B	C	D	E	F	G	H
1					**Sample data: ADLIB**			
2								
3	***Name >***	***Parent***	***Level***	***Ref no***	***Title***	***Extent***	***year from***	***year to***
4	***Adlib Tag >***	***bt***	***gv***	***IN***	***TI***	***DA***	***DS***	***DE***
5	***Column no >***	***1***	***2***	***3***	***4***	***5***	***6***	***7***
6								
7	***First import row >***		**Fonds**	**D275**	**Stroud Cooperative Society**		**1899**	**1991**
8		**D275**	**Sub-Fonds**	**D275A**	**Stroud Cooperative Society: Governance**		**1899**	**1991**
9		**D275A**	**Series**	**D275A/1**	**Stroud Cooperative Society: Rules**		**1899**	**1991**
10		D275A/1	Item	D275A/1/1	Stroud Cooperative Society Rules	1 document	1899	1975
11		D275A/1	Item	D275A/1/2	Stroud Cooperative Society Rules	1 document	1908	1991
12		D275A/1	Item	D275A/1/3	Stroud Cooperative Society Rules	2 documents	1917	1987
13		D275A/1	Item	D275A/1/4	Stroud Cooperative Society Partial Amendment	1 document	1990	1991
14		D275A/1	Item	D275A/1/5	Stroud Cooperative Society Partial Amendment	1 document	1990	1991
15		D275A/1	Item	D275A/1/6	Stroud Cooperative Society Partial Amendment	1 document	1929	1980
16		**D275A**	**Series**	**D275A/2**	**Stroud Cooperative Society: Reports and Balance Sheets**		**1920**	**1970**
17		D275A/2	Item	D275A/2/1	Stroud Cooperative Society Quarterly Report and Balance Sheet	3 documents	1920	1936
18		D275A/2	Item	D275A/2/2	Stroud Cooperative Society Half Yearly Reports and Balance Sheets	12 documents	1930	1959
19		D275A/2	Item	D275A/2/3	Stroud Cooperative Society Draft Balance sheet	1 document	1943	1950
20		D275A/2	Item	D275A/2/4	Stroud Cooperative Society Details of Wages Costs Per Pound of Sales	1 document	1951	1960

<< END OF THIS UNIT >>

UNIT P5

Migration between Excel and specialist systems: **THE ARCHIVES HUB**

[Many thanks to the team at The Archives Hub for checking and revising this unit, December 2018]

Excel- compatible transfer formats

- Encoded Archival Description (EAD)

EAD is a special case of XML (see UNITS N3 and N4)

Location of Import facility
https://editor.archiveshub.jisc.ac.uk/
Note that you need to log in to use the EAD editor. Accounts are free, but you need to contact the Archives Hub team at contributors.hub@jisc.ac.uk so that they can create an account for you.

Documentation
- https://archiveshub.jisc.ac.uk/contributing/
- https://archiveshub.jisc.ac.uk/eadeditor/
- https://archiveshub.jisc.ac.uk/eadforthehub/

Overview of import procedure
- Ensure that your data for import is in a *.xml document, compliant with the HUB EAD requirements [see below for summary of one method for preparing data for import]
- Open the Archives Hub EAD editor
- In the top menu bar, choose 'Upload description' and select the xml file to be uploaded from your computer (see screenshot 1)
- Click 'Upload'
- The Hub software checks and uploads the data, which is displayed within the Hub editor, where it can be manually checked and edited (see screenshot 2)
- The Hub Editor is strict in its requirements, and if your xml document is non-conformant in any way you may find the upload fails and you receive an error message. If this happens feel free to contact the Hub team at contributors.hub@jisc.ac.uk for help, including the file and the error message if possible.

Excel format and structure requirements
- Defining the archival hierarchy: Nested components within a Collection Level record.
- XML document, compliant with the Hub's EAD standard (see https://archiveshub.jisc.ac.uk/eadforthehub/

Export from the Archives Hub

- Open the Hub EAD Editor
- In the top menu choose 'Edit description', and then open the collection which you wish to export
- Choose one of the options: 'Download' will download a copy of the xml file to your computer; 'Email' will allow you to email the xml file to yourself; 'EAD' will open the xml file in a new browser tab (see screenshot 2)
- The file is saved as a *. XML file which you can view and edit in Excel or in a text editor

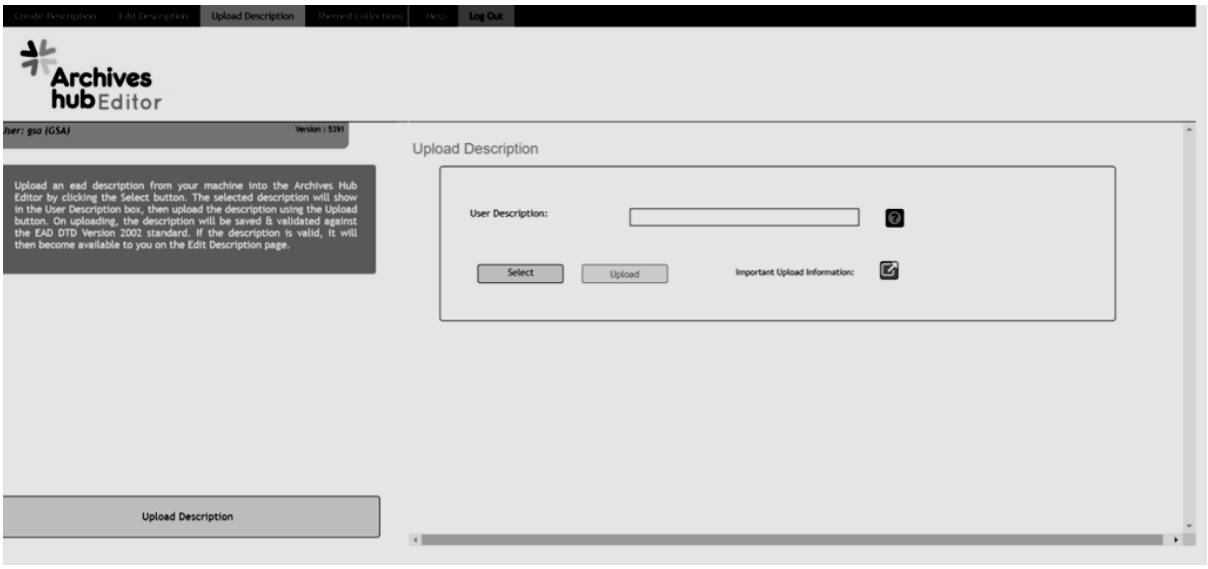
Summary of a method of preparing data for import into the Archives Hub

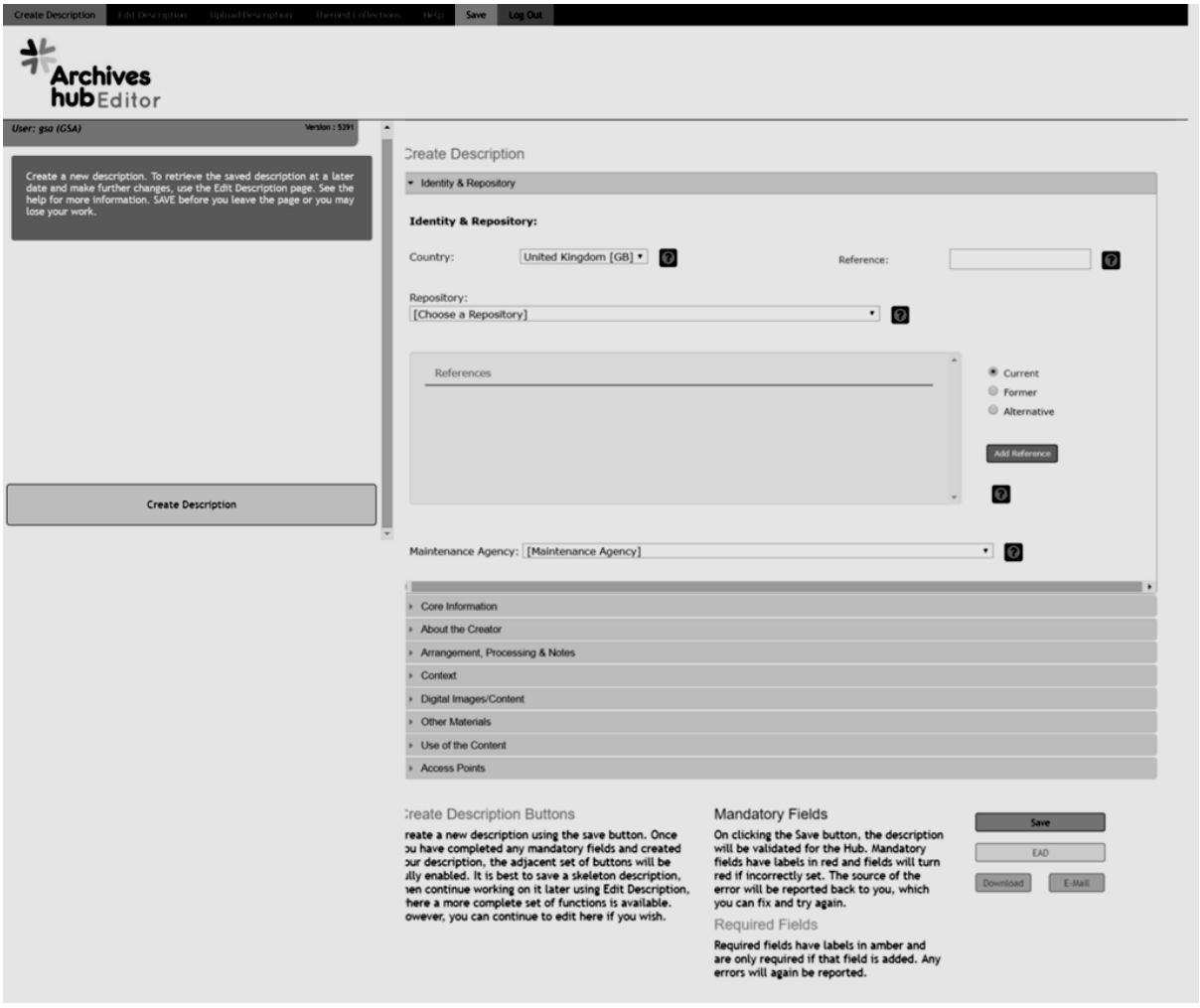
One way to transform an Excel catalogue into an EAD document which can be uploaded to the Archives Hub is to create an outline catalogue using the Hub EAD Editor. This is exported and serves as a template for the complete set of data, which is mapped to the template, edited in a text editor and then uploaded to the Hub. However this is fairly tricky, requiring use of a text editor and good understanding of EAD).

The following is a summary of the method: a much more detailed step by step description and a set of example files is available. This method is by no means definitive, and there may be a simpler way to achieve the same result.

- Use the Hub editor to create an outline catalogue with two component records, all treated as component 1 (that is, with no nesting of components), and export the outline catalogue as an XML document.
- Open the XML document in a text editor. Add new (temporary) tag pairs which will be used to define the true component nesting (for example, following the closing tag </c01> with the new tag pair <endcomponent> *a* </endcomponent>)
- Use the edited document as a template to match the catalogue data columns in Excel, and as an xml map: the columns represented by the new (temporary) tag pairs are used to record the component positions.
- Open the resulting xml document in a text editor, and using find and replace change the temporary tag pairs and the original component tag to the correct components. For example, change <c01><startcomponent> *item*</startcomponent> to <c03 level="item">, or change <endcomponent> *end series and subseries and item*</endcomponent></c01> to the three closing tags </c03> </c02> </c01>

ARCHIVES HUB SCREEN SHOTS





<< END OF THIS UNIT >>