UNIVERSITY OF LEICESTER

School of History

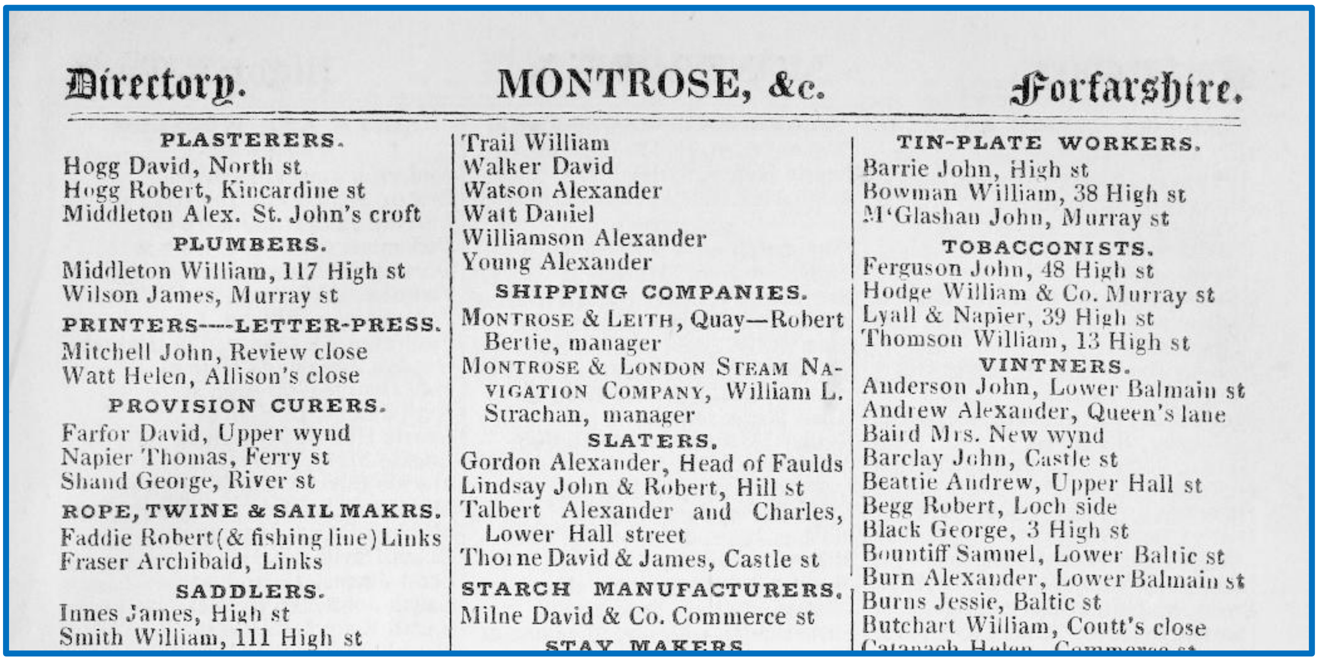
www2.le.ac.uk/departments/historical/outreach

Counting sources

As mentioned in the 'Approaches' section above, there are many historians who take a 'quantitative' approach to history. This essentially involves the numerical analysis of historical evidence. This might involve information that was collected in the past, such as financial data from company accounts. Or it might involve counting up sources themselves – the number of bastardy bonds from a region might, for instance, be taken as an indicator of the level of extra-marital childbirth.

Undertaking quantitative analysis is not as simple as putting data into a spreadsheet and making a few calculations. We need to decide what information we are going to use, how we are going to record it, how we will analyse it, and how we will analyse these results.

Take, for instance, the following source. It looks like there's lots of information here. But how would we go about entering this information into a spreadsheet or a database?



Directory. MONTROSE, &c. Forfarſhire.

PLASTERERS.
Hogg David, North st
Hogg Robert, Kincardine st
Middleton Alex. St. John's croft

PLUMBERS.
Middleton William, 117 High st
Wilson James, Murray st

PRINTERS----LETTER-PRESS.
Mitchell John, Review close
Watt Helen, Allison's close

PROVISION CURERS.
Farfor David, Upper wynd
Napier Thomas, Ferry st
Shand George, River st

ROPE, TWINE & SAIL MAKRS.
Faddie Robert (& fishing line) Links
Fraser Archibald, Links

SADDLERS.
Innes James, High st
Smith William, 111 High st

Trail William
Walker David
Watson Alexander
Watt Daniel
Williamson Alexander
Young Alexander

SHIPPING COMPANIES.
Montrose & Leith, Quay—Robert Bertie, manager
Montrose & London Steam Navigation Company, William L. Strachan, manager

SLATERS.
Gordon Alexander, Head of Faulds
Lindsay John & Robert, Hill st
Talbert Alexander and Charles, Lower Hall street
Thorne David & James, Castle st

STARCH MANUFACTURERS.
Milne David & Co. Commerce st

STAY MAKERS.

TIN-PLATE WORKERS.
Barrie John, High st
Bowman William, 38 High st
M'Glashan John, Murray st

TOBACCONISTS.
Ferguson John, 48 High st
Hodge William & Co. Murray st
Lyall & Napier, 39 High st
Thomson William, 13 High st

VINTNERS.
Anderson John, Lower Balmain st
Andrew Alexander, Queen's lane
Baird Mrs. New wynd
Barclay John, Castle st
Beattie Andrew, Upper Hall st
Begg Robert, Loch side
Black George, 3 High st
Bountiff Samuel, Lower Baltic st
Burn Alexander, Lower Balmain st
Burns Jessie, Baltic st
Butchart William, Coutt's close
Catanach Helen, Commerce st

Figure 1: Pigot and Co.'s National Commercial Directory for the whole of Scotland and of the Isle of Man

Published online by the University
of Leicester, 2014

Arts & Humanities
Research Council

CONNECTED
COMMUNITIES

Dates

If we go further back in history, even things like the date can be problematic. Before 1752 Britain used the Julian calendar. In September of that year, 11 days were skipped, as the country joined Europe in using the Gregorian system. Before this date, the start of the year was sometimes taken to be 25 March. A date such as ‘1 February 1722’ might, for instance, actually be in one of two years. If you were making a database in which you calculated people’s ages from baptism records, how would you deal with this?

(1837), http://archive.org/details/pigotcosnational1837dire . Please note that this image is not covered by the same Creative Commons licence as this document. Follow the link for details of the licence.

Let’s say we are interested in the history of small businesses in this town. How are we going to group these professions? Do we group the bakers and pastry makers together? What about the iron workers and the blacksmiths? If we keep everything separate, then how can we see the bigger picture?

One problem is that sometimes the categories that we use can change. Streets can be renamed, boundaries expanded, or the coverage of a source such as a trade directory can change. How, for instance, would we go about comparing a directory from Leicester in 1880 with that from 1900?

The simplest advice is simply to be consistent. You might resolve simply to record occupations as they appear in your trade directory, or dates as they appear in your baptismal register. Similarly, you might resolve to use the parish boundaries of the 1850s for your geographic analysis, and stick with these throughout.

This may be something that we wish to carry only so far, however. In the seventeenth and eighteenth centuries, for instance, it was common for names to spelled in an extremely inconsistent way. If we record these names in their own unique fashion every time we see them, do we not risk greatly inflating the population of a parish by counting John Smith, John Smyth and John Smithe as three separate people?

A sound piece of advice is to take a record of any decisions that you take regarding these issues. You could do this in a separate document, or include a column in your database or spreadsheet in which you note down any decisions which you think were important. This will at least allow you to justify what you have done in future.

Analysing data

As mentioned, consistency is important when conducting quantitative analysis. But the ways that you choose to analyse the data can also be important. Take, for instance, the following table. What would be the best way of working out the typical age of a sailor in the Victorian navy?

Name	Age
Richard Smith	20
Stephen Thomas	22
John Walters	58
Thomas Turner	24
William Jones	19

When people talk about 'the average' in everyday life, what they are usually referring to is the mean average. This involves adding all the different items, and then dividing them by the number of items. So in this case, we would perform the following sum:

Average = (20 + 22 + 58 + 24 +19) ÷ 5 = 28.6

But is it really fair to say that the average age of these sailors is 28.6? It's clear that the age of John Walters in the table above has greatly skewed the results.

A different method of calculating an average might, therefore, be necessary. There are two other options:

- Median: the middle value of a series (or the mean of the two middle values if the series has no single middle value). In the above example, there series is as follows: 19, 20, 22, 24, 58. The median age is, therefore, 22.
- Mode: the value that occurs most often. In the following sequence of numbers, for instance, the mode is five: 1, 1, 2, 3, 3, 4, 5, 5, 5, 6, 6, 7, 8, 9, 9. If we applied this approach to the table above, we'd find that there is no mode, since all of the ages are unique.

From all this we can see that the approach we take can have a profound effect on our results. Looking at the table above, which do you think is the best way to characterise the average ages: 28.6, 22, or no discernible pattern? In this case, the median age – 22 – intuitively looks most useful.

This illustrates how quantitative analysis is not an inherently ‘neutral’ activity. We make decisions about how to record and analyse information, and this can have a great influence on our results. The following sections go into more detail about some of the common issues that historians encounter. But there may be some cases in which you have a specific problem, for which more detailed help is needed. For this reason, you may find it useful to consult the following resource, which go into considerably more detail on conducting quantitative historical research.

- Designing databases for historical research (free course), a free online course from the Institute of Historical Research, University of London, accessible at http://www.history.ac.uk/research-training/courses/designing-databases

Data relationships

In a section above, we discussed the difficulty of classifying occupations, and the challenge that is posed by inconsistent spelling of people’s names. Fortunately there are ways that we can avoid having to make an either/or decision when we put together our data. You might, for instance, try to include unique job titles in your database, while also linking these to a standardised set of occupations.

If you are simply creating a table of data in a spreadsheet program such as Microsoft Excel or LibreOffice Calc, then the following example layout might be sufficient.

Name	Job title	Occupation Category
John Smith	Builder’s apprentice	Construction
Jane Blackwell	Seamstress	Textiles
James Taylor	Mason	Construction
Stephen Jones	Lawyer	Professional
Hannah Smyth	Wool carder	Textiles
George Way	Tailor	Textiles

This approach is helpful because it allows the best of both words—you can preserve the original information, but also put it into categories.

Do you agree with the example categories I’ve chosen here? One might argue that grouping a wool carder –which involves working with unprocessed wool – with a tailor is a questionable decision. Tailors were skilled workers who produced finished goods from a range of materials. Perhaps grouping the tailor with the mason would be more appropriate? There is no simple answer to this question. This is why it’s important to consider how you’re using the data, what you want to find out, and to maintain a record of the decisions that you take.

Linking data together

If you’re using a more advanced program such as Microsoft Access, it’s possible to link different sets of data.

Let’s say that you were editing census data and produced the following table.

Name	Age	Street	Town	County
John Smith	17	Carver Street	Sheffield	Yorkshire
Jane Blackwell	22	Carver Street	Sheffield	Yorkshire
James Taylor	31	Garden Lane	Sheffield	Yorkshire
Stephen Jones	42	King Square	Sheffield	Yorkshire
Hannah Smyth	25	Turner Row	Sheffield	Yorkshire

Those ‘town’ and ‘county’ columns involve huge amounts of repetition. They also mean that if you want to change the county from ‘Yorkshire’ to ‘South Yorkshire’, you’d have to replace every value. Similarly, if you decided that actually we need to add parishes to the data, you would have to go through the entire table again.

The solution to this problem is to create separate tables and use shared variables to establish a relationship between them.

In our first table, for instance, we only need to include the first three columns in the table above. Our second table could include a list of street names. Each street name could specify the city. A third table could then comprise a list of towns in Britain, with the county in which they are located. You will see from the next figure, that some of these tables share the same data:

Table 1: Individuals

Name	*Age*	*Street*
John Smith	17	Carver Street
Jane Blackwell	22	Carver Street
James Taylor	31	Garden Lane
Stephen Jones	42	King Square
Hannah Smyth	25	Turner Row

Table 2: Sheffield streets

Street	*Town*
Carver Street	Sheffield
King Square	Sheffield
Turner Row	Sheffield
Garden Lane	Sheffield

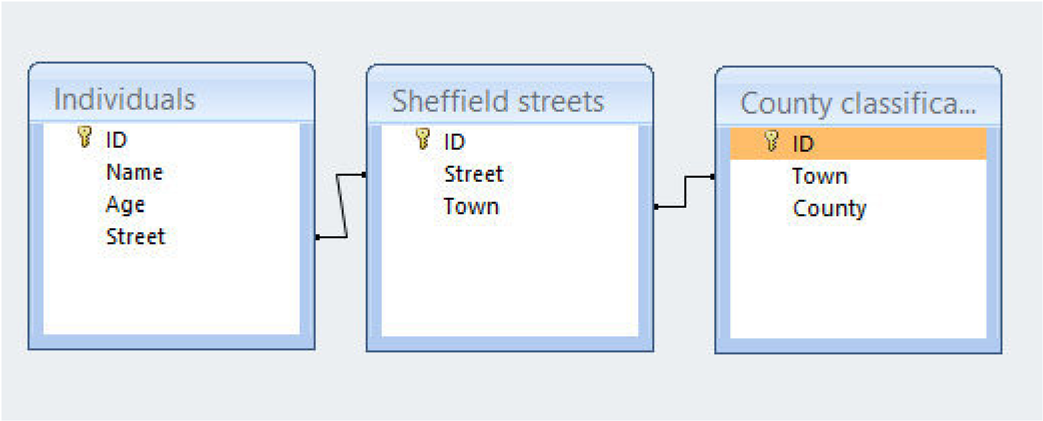
Table 3: County classifications

Town	*County*
Sheffield	Yorkshire
Stoke	Staffordshire
Buxton	Derbyshire
Chester	Chesire
Leeds	Yorkshire

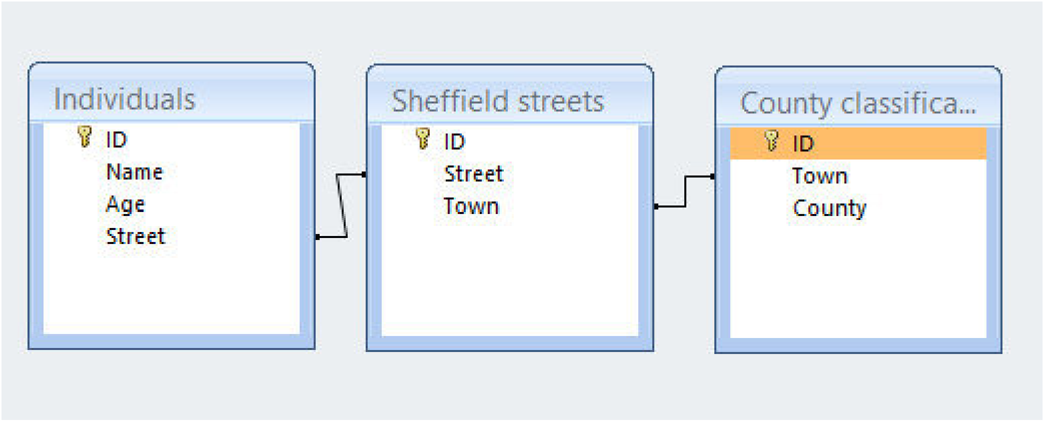
If we really wanted to, we could follow this method to its conclusion, and link individuals with countries and even continents!

The advantage of this method is that if you wanted to assign the streets to parishes, you could simply add a ‘Parishes’ column to Table 2, and define which parish each street was in. This would avoid all the repetition associated with adding parish data to every record in Table 1.

Database software such as Microsoft Access will enable you to link these tables as follows:



This allows you to conduct queries which take these links into account. You might want to calculate the average age of people in Sheffield, for instance. You can do this by creating a query which groups the data according to the ‘Town’ column in Table 3, and then calculates the average of the ‘Age’ column in Table 1. It would look as follows:



The other advantage of structuring your data in this way is that you may be able to create links between entirely different datasets. In the example above, for instance, we have included occupation data. If we had another dataset which included occupation data, such as a trade directory, then it might be possible to link these two items.