

CREATING LIFE HISTORIES AND FAMILY TREES FROM NINETEENTH-CENTURY CENSUS RECORDS, PARISH REGISTERS AND OTHER SOURCES

Peter Tilley

Peter Tilley is currently at the London School of Economics and Political Science working towards an M. Phil./Ph. D. on internal migration in Victorian Britain. He was previously Honorary Research Fellow at Kingston University, acting as Project Manager for the Kingston Local History Project. Prior to that he spent many years in the computer software industry.

Introduction

Since early 1996 the Kingston Local History Project, based at the Centre for Local History Studies at Kingston University, has been building a computer database from the variety of written and pictorial resources available in the local community. This will be used to explore various aspects of Kingston's socio-economic history in the late-nineteenth century. For two years the project had no budget, so the early work was undertaken by an unpaid project manager (the author) and volunteers recruited by him from the local community. In 1997 limited funds became available from the university to enable the employment of a part time administration assistant followed two years later by funding for a database administrator.¹ This paper is primarily focussed on a description of the process involved in producing personal life histories.

In line with the thinking of J.D. Marshall, the project has encouraged the active participation of the local community in collecting and recording relevant data.² It also makes the resulting life histories and other records available to genealogists, local historians and other interested academics through a user-friendly version of the database at the local history room of the Royal Borough of Kingston's Heritage Centre. Whilst the old market town of Kingston-on-Thames is the main area for study, the village of Blechingley, some 20 miles to the south, is also part of the project, mainly as a pilot study but also as a rural alternative to the urbanised community of Kingston.

Up to the mid-nineteenth century, Kingston had been a market town and the first stop for stage coaches leaving London for Portsmouth and the West Country. However, the coaching trade went into terminal decline when the railways arrived in the area in 1837. Although on the main railway line, for a variety of reasons Kingston station was located about one and a half miles outside of the town in the small village of Surbiton. A local property

entrepreneur saw the potential of a green-field site on the river Thames only 40 minutes away from London by train and set about building spacious middle class villas together with the necessary supporting artisan cottages. The success of his enterprise can be judged by the 353 per cent population growth for Kingston between 1841 and 1891 when the national population only grew by 82 per cent. Throughout this period, Kingston had no significant industries either within the town or in close proximity. Indeed, based on the Booth/Armstrong classification system, the three occupational sectors showing substantial growth rates were 'dealing' (shops), 'public service and professional', and 'domestic service', whilst agriculture and manufacturing were declining.³ Such a profile could be expected from the middle class suburb that Kingston was becoming.

Blechingley is situated on the main road from Guildford and Reigate to Canterbury and the channel ports on the Kent coast. With a population in this period averaging about 1,700 it could be considered a fairly large village but its overall growth was almost zero. Whilst the agricultural depression of the 1870s may have accounted for some of this lack of growth the fact that the nearest railway station was some four miles away in the village of Nutfield could have inhibited any influx of middle class commuters. The population of Nutfield itself grew by 90 per cent between 1841 and 1891 and the bulk of this growth was in the final 30 years.

Because the work on Kingston is ongoing and incomplete, this paper will focus on complete life histories and family trees from Blechingley, but will describe a method common to both locations. Figures 1-3 illustrate the types of report that can be produced. The benefits to genealogists and family historians of such 'on demand' life histories and their associated family trees are obvious, but the underlying data are also available for analysis. When up to six census records are available for any individual, age reporting can be easily checked for accuracy, as can place of birth. Individual occupational changes over time can be easily monitored against national trends and the commonality of occupations for fathers and sons can also be checked, confirming the more limited data from marriage records. The factors affecting social mobility within a community may well be revealed and models for migration behaviour beyond the community can be built.

Previously, most family reconstitution has been based on a period prior to the late-nineteenth century, using parish registers to derive statistics on nuptiality, fertility and mortality for that community.⁴ The additional data available in life histories enables the researcher to compile much more comprehensive family structures. Eckstein and Hinde recently reported in this journal on a study of family reconstitution for the period 1841-1891 using parish registers and the census enumerators' books.⁵ With very little modification their methods could make use of life history records as described in this paper. Life history data would also appear to be an ideal source for event history analysis.⁶

Source data

At this stage, four main sources of data have been used to construct the life histories. First are the census enumerators' books (CEBs) for 1851 to 1891. Although there were slight variations of content over the years, for the purposes of this paper each record in the CEBs can be thought of as consisting of at least name, address, relationship to head of household, marital status, age, sex, occupation, place of birth (town or village and county, plus country if not England), and a statement of certain disabilities.⁷ Second, we have parish registers of baptisms, marriages and burials for 1840 to 1900. From 1813 parish registers for the established church had a fixed columnar format. Baptism records contained the date of baptism, the child's forename and surname, the father's and mother's names, the father's trade or profession, the parents' abode, and the name of the officiating minister. In some parishes the date of birth is also given, albeit intermittently. Marriage registers showed the date of the marriage, and for both bride and groom: name, marital status, age (if over 21 then this would usually state 'full' or 'of age'), occupation, abode, father's name and father's occupation. Bride, groom and witnesses also signed the register or left their mark and the officiating minister signed his name. Burial registers were quite simple, showing only the date of the burial; the name, age at death, and abode of the deceased, and the officiating minister.

Third, the parish burial registers are augmented by cemetery records for burials after 1855. Municipal cemeteries were introduced in the 1850s when many of the parish churches were having difficulty in finding space in their own graveyards. In Kingston all denominations were accepted in the cemetery, although different religions were often allocated different areas. The cemetery burial register records are of good quality and contain additional information beyond the parish burial registers such as grave location, type of grave and whether the ground was unconsecrated. In Blechingley they also contain the depth of the grave and the fee paid. Finally, there are trade and street directories published by the likes of Kellys. These tended to contain records of two types of people: those who wished to advertise their trade, such as grocers and carpenters, and those who wished to advertise their perceived social status by virtue of an entry in the directory. Listings were usually in street sequence as well as trade sequence where appropriate.⁸ Later additions could include the 1841 CEBs (already in for Blechingley), electoral rolls, rate books, school registers, workhouse admission records, newspaper articles and indeed any source with local nominal data.

With the exception of the 1881 census where a machine-readable copy was provided by the Genealogical Society of Utah (Mormons), all this source data needed to be keyed into computers. It was decided that, where possible, hard copies would be obtained so that data entry could be undertaken by our team of local volunteers (currently numbering about 30).⁹ These range in age from 25 to 81 and have proved to be a very valuable asset. Most work at home on their own computers using software and instruction manuals prepared by the centre; some come to the centre

Table 1 Volume of source data for Kingston and Blechingley

Source	Number of records	
	Kingston	Blechingley
Census enumerators' books	140,000	11,000
Baptism registers	13,800	2,900
Marriage registers	10,300	500
Burial registers	33,065	2,500
Local directories	See note below	730

Notes: The baptism registers for Kingston are incomplete. The census enumerators' books for Blechingley include the 1841 census. The burial registers for Blechingley go up to 1917. The local directories for Kingston have not yet been incorporated into the database. There are three publishers of local directories in Kingston during this period and as yet the books are not in electronic form. Current effort is on consolidating census records and parish registers, with local directories scheduled as one of the next projects. Because they are printed we have been considering using optical character recognition on scanned images but the problems involved may well outweigh the benefits.

twice a week for checking and some do both. All the Blechingley data were keyed in by the 81-year old. The volumes involved for both locations are shown in Table 1.

The keyed-in data were imported to the main database, where the content was checked and validated. Checking was a visual process from print-outs but the computer was able to carry out validity checks which revealed, for instance, a boy called Mary, a female blacksmith and a woman of 75 claiming to be mother of a five year old. To ensure source integrity, wherever possible we have retained the original written information so the archive section of the database may contain apparent anomalies such as the examples above. Of course, the analysis section can be adjusted to reflect a more accurate view of the source data.

Database design features

The database package used for the project is Microsoft Access, which is part of the Microsoft Office Professional suite of applications. Within Access, *tables* store the data, *queries* sort, filter and analyse data in the tables, *forms* provide a user-friendly interaction to processing, and *reports* facilitate the design and formatting of printed output. To ensure portability, most processing for the project uses only these standard features within Access: hence users do not require a knowledge of programming. A new low-specification personal computer would have sufficient power and storage capacity to handle many times the volumes required.

Figure 1 Generation chart for Sargant family

65573	William SARGANT b.1808 (Spouse: Harriett) d.1879
66820	William T SARGANT b.1831 d.1866
66821	Mary SARGANT b.1834 d.1854
66822	Alice SARGANT b.1845 d.1896
66823	Eliza SARGANT b.1846 d.1894
66824	Alfred Sneed SARGANT b.1848 m.1880 (Spouse: Maria) d.1917
71623	Irene K SARGANT b.1882
71624	Violet SARGANT b.1883
74622	Gladys Reine SARGANT b.1885 d.1885
74662	Alfred Norman SARGANT b.1886 d.1888
71625	William G SARGANT b.1890
74853	Dudley Howard SARGANT b.1894 d.1898
66825	Edward SARGANT b.1850 d.1868
73924	Herbert SARGANT b.1852 d.1853
68344	Joseph SARGANT b.1855
68043	Edith SARGANT b.1858 m.1881
68345	Marian SARGANT b.1860 d.1883
68997	Ada SARGANT b.1864 m.1884 (Spouse: Alfred)

When using Access, data are stored as *fields* within *records*. A table is a collection of records of the same type. Within data tables, each record can be split conceptually into two sections. The archive section contains the data as recorded in the original documents. The other section is made up of added fields containing the coded and standardised data needed for analysis and record linkage. Each record in every table also contains a field that gives it a unique identity (UID) within the table.

After the record linkage routines have been completed, the UIDs can be grouped to identify individuals in the various data tables. For the Kingston project the table that holds these groupings is called the *person table*, since it has a uniquely identified record (a *person record*) for each known individual. By definition, an individual can have only one record per census year, one baptism record and one burial record. The relevant records pertaining to that person are identified by their UID as stored in the person record (their *person ID*). This technique is not appropriate for marriage records, local directories and parent fields within baptisms and marriages where multiple records can be linked to the same individual. To cater for these possibilities the technique is to store the UID of the relevant person record as a field in the appropriate record within the data table. This method is known as using a 'foreign key' and for record retrieval it is less efficient than the direct key method

previously described. However for the application described and with the processing speed of modern computers such concerns over efficiency are academic.

When the original data is unsuitable for analysis or record linkage it has to be converted to a standardised or coded version. This is achieved with the use of *lookup tables*. Records in lookup tables have two fields, one for all possible variations within the original data and one for the appropriate standard version. By matching the original data with the first field of the lookup table, the second field containing the standardised or coded version is available for processing. This method is used to standardise forename, county of birth, relationship to head of household, and marital status. It also provides codes for occupations based on the Booth/Armstrong classification.¹⁰

Record linkage

The objective of record linkage is to identify the same person in different source material.¹¹ For Kingston life histories, that source material will be census records of different years, parish registers of baptisms, marriages and burials, cemetery records and entries in street/trade directories. By early 1997, the database contained records for the censuses of 1861 and 1871 and it became necessary to devise routines for inter-censal linking. The volume of data involved was too large to consider purely manual methods but a search of the literature revealed several papers and book chapters involving automated record linkage. Some early examples from the 1960s and 1970s (mostly concerned with seventeenth- and eighteenth-century parish registers) used large main frame computers and sophisticated programming routines, neither of which are applicable to computer processing in 1997.¹² Other previous examples involved family based record linkage for North American census data, computer-aided linkage for eighteenth century Cheshire poll books and automated linkage between CEBs and civil register data for mid-Wales.¹³ On examination none of the methods they used was felt to be wholly appropriate to the exercise in linkage which we were wishing to undertake. The study of mid-Wales by H.R. Davies had most relevance, but his procedures made no use of family context and were fully automated using a software package which is no longer readily available.¹⁴ The North American study described attempts to incorporate a family context but lacked sophistication in the use of computer software, having been written in the 1980s.

We were tempted by the multiple pass linkage algorithms described in a recent paper by Harvey, Green and Corfield.¹⁵ These constitute a fully automated system that uses successive passes through the data to arrive at linkages with confidence levels attached. When used by them on the Westminster poll books of 1784 and 1788 it produced impressive results. We set about coding their concept as Access routines for the CEBs and completed this in two days. Our results appeared equally impressive in that a 39 per cent linkage rate was achieved with a confidence level exceeding 80 per cent whilst

a 50 per cent linkage rate could be achieved if the confidence level was lowered to 28 per cent. Before accepting these results we decided to do some 'spot-testing', and it was at this point that we began to have doubts about the technique. Of a random sample of the matches found by the algorithms, some proved obviously false when checked against the original manuscript CEBs. Furthermore, some links which we had identified as true during preliminary experiments were not made by the application of these algorithms.

We decided to examine each potential match derived by the application of the multiple pass algorithms in the context of the household. To undertake this as a purely manual exercise would have been prohibitively demanding on time, so we looked at using facilities within the database. An Access form was designed which displayed the two records selected as potential matches. The two 'individuals' referred to by the records on the form are usually members of larger families or households and it is possible that information about this could help to confirm or deny the match. For example, if the two 'individuals' shown on the form have different parents in two census years the potential match is likely to be false. Two sub-forms were used to display the household members associated with the two entries on the main form. The reviewer indicated his or her decision using appropriate *match flags* on the main form, the three options being 'review', 'false' and 'true', with 'initial' as the unallocated setting (these are shown towards the top right-hand corner of Figure 4). Using his or her judgement, the researcher was required to allocate either 'true' or 'false' to each potential match. If he or she wished to consult the primary sources, or to discuss the potential match with colleagues the form was marked as 'review'.

The conclusions of this exercise confirmed our doubts. Over 3,600 (or 44 per cent) of the matches found by the automatic routines proved to be false when considered in context of the household, whilst a further 1,900 true matches were missed by the automatic linkage routines.¹⁶

However, the effectiveness and speed of the manual review process seemed to offer a way forward so we set about associating it with an automated system of extracting potential links through the use of algorithms. We were looking for algorithms which would create these potential links by giving maximum exposure to the data whilst minimising the number of false matches. The practicalities of various algorithms are illustrated in Table 2 which shows for 1871–1881 the number of additional true links tailing off for the 'weaker' algorithms whilst the number of potential links to be reviewed escalates up to 150,000. After much experimentation we have come up with a set of five algorithmic approaches which would appear to offer a pragmatic and effective approach. These have been used on five sets of census data for Kingston, six sets for the villages of Blechingley and Nutfield and three sets for the industrial town of Middlesbrough. In all cases we have been able to establish a linkage rate of between 35 and 40 percent of the population with what we would like to think of as 100 per cent confidence.

Table 2 Effectiveness of various record linkage algorithms for Kingston 1871–1881

Algorithm	Potential matches	True matches
Sex, Surname, Standard Forename, Year of Birth, Standard County	4,283	4,128
Sex, Surname, Standard Forename, Standard County, Year of Birth \pm 5	7,306	6,318
Sex, Soundex Surname Code, Short Standard Forename, Standard County, Year of Birth \pm 5	13,544	7,765
Sex, Soundex Surname Code, Short Standard Forename, Standard County	47,759	8,117
Sex, Soundex Surname Code, Standard Forename,	60,371	8,319
Sex, Soundex Surname Code, Short Standard Forename	150,360	8,889
Combination of algorithms and other methods e.g. Pattern matching, Family forename matching	19,604	10,194

The match criteria required within the five algorithmic approaches are described below:

1. Matches on Short Standard Forename, Soundex Surname Code, Sex, Standard County of Birth, and Year of Birth (calculated by subtracting recorded age from the census year) within five years.¹⁷
2. Matches on Short Standard Forename, Full Surname, Sex, and Year of Birth within five years. Compared with algorithm 1, this algorithm has been weakened by removing County of Birth but tightened slightly by using full surname. Whilst it will produce many more potential matches it is likely to have a low hit rate for true matches. It is aimed at catching mis-recorded or missing county of birth.
3. Matches on Short Standard Forename, Full Surname, Sex, and Standard County of Birth. Aimed at mis-recorded age, this will tend to create a large number of potential matches, though this number can be reduced if Year of Birth is included with a tolerance of 20 years. This latter measure avoids matching a child with a parent of the same forename.
4. Matches on Short Standard Forename, Sex, Standard County of Birth, Year of Birth within five years, and Surnames with a suitable Guth matching score.¹⁸
5. Matches on Short Standard Forename, Sex, Standard County of Birth, Year of Birth within five years, plus two or more other family members having the same personal details in each census year.

Algorithm 1 is the basic algorithm and should produce about 75 per cent of the

true matches. Algorithms 2–5 are likely to produce another 15 per cent in total. The remaining 10 per cent are identified without algorithms.

When reviewing a potential match which has been produced by one of the algorithms it may be obvious to the researcher that other records in the household should be matched. Consider Figure 4, which shows the review of the potential link for Louisa Campbell. This potential link was identified by algorithm 1 and is clearly true. The link between the two records for her daughter Alice will need to be identified later by algorithm 2 because place of birth is missing in 1891. However, this could be pre-empted by the researcher. The field labelled F1 is blank if that record has not been identified by an algorithm so if an obvious match within the sub-forms is displaying a pair of blank F1 fields then that potential link can be introduced for consideration at a later phase. The researcher keys the record identifier from the second table (shown as field F2 in the second sub-form) into the blank field F2 on the first sub-form so updating the first table. Moreover, looking at the whole household it would appear obvious that all three people in 1881 are in fact still in the same household in 1891 and the field F2 should also be updated for the servant named in 1881 as Elizabeth Campbell. An experienced researcher will be aware of the possibility that in 1881 the enumerator carelessly wrote 'do' for all surnames subsequent to the head. This could explain why Elizabeth Perrott was previously recorded with the same surname as the head of the household. The absence of place and county of birth for Elizabeth in 1891 would weaken further the potential for automatic linkage.

A pseudo-algorithm extracts records identified by a completed F2 and adds them to the link table for review by the researcher. The example in Figure 5 shows two entries for a William Brockwell for 1871 and 1881. In isolation, most fully automated routines would assess that all relevant fields match perfectly and assign a true result with a confidence level of 100 per cent. In reality, the household detail revealed in the sub-forms shows that the two families are completely different and the match should be false. Any record linkage procedure should be able to cope with these kinds of scenarios.

From the above it can be seen that we think it is impossible to define in advance rules for researchers to follow when examining possible links. Hence they are told to allocate the match flag 'true' only if they are sure in their own mind that the two records relate to the same person. 'False' is to be allocated for any cases where there is uncertainty unless the researcher feels that possibly the uncertainty could be resolved by looking at the source records or by discussing it with colleagues. In those cases the allocation is 'review'. This whole approach is at odds with the method usual for fully automated routines where the true decision is a probability statement based on the best fit amongst competing options.

There are several safeguards built into our procedures. All potential links are maintained in a *link table* together with the researcher's assessment of true, false or review. This enables records from either table which have been previously classified as part of a true pair to be excluded from future

Figure 4 Example of census record linkage (true)

Kingston - Review Record Linkage

E D/Page/Line: 03 | 0015 | 21 | 1881 | 04 | 0017 | 27 | 1891

Forename: Louisa | Louisa
Surname: CAMPBELL | Campbell
Rel Head/Status: Head | W | Head | Widow
Address: 2 Tudor V Tudor Rd | Tudor Road
Parish of Birth: --- | Parliman Square
County of Birth: --- | London
Occupation: Annuitant | Living on own means
Age/YOB/Booth: 45 | 1836 | | | 55 | 1836 | | |

Match Status: True False Review Initial
Sort Seq: ED Seq, Name, Match/ED, Match/Name
EXIT

1881										
F2	F1	Forename	Surname	Relation	Marita	Age	Y O B	Town Born	County	Occupation
	01	Louisa	CAMPBELL	Head	W	45	1836	---	Middlesex	Annuitant
		Alice R.	CAMPBELL	Daur	-	5	1876	St Heliers	Jersey	Scholar
		Elizabeth M.	CAMPBELL	Serv	Unm	29	1852	St Aubins	Jersey	General Domestic Serv

1891										
F2	F1	Forename	Surname	Relation	Marita	Age	Y O B	Town Born	County	Occupation
35972	01	Louisa	Campbell	Head	Widow	55	1836	Parliman Squa	Middlesex	Living on own means
35973		Alice H	Campbell	Daur	S	15	1876			Scholar
35974		Elizabeth M	Perrott	Serv	S	39	1852			General servant
35975		Mabel	PECKHAM	Visitor		19	1872			Scholar

Record: 13430 of 24767

algorithms. At the end of the review of a particular algorithm run, the system also checks whether any records have been classified as part of a true linkage more than once and, if such records are found, it changes the match flags in the link table back to 'initial' for further review of the offending linkages.

On census records our experience has shown that using these five algorithms the number of potential matches can be estimated at 75 per cent of the total population in the smaller of the two census tables. Whilst not essential, it is preferable to assess all potential links from the first algorithm before applying the subsequent algorithms. This prevents any records which are part of a true match from being considered by the other algorithms. As a guide to the required time we have found that a competent researcher can assess 200 potential matches per hour. Hence the processing time for the linkage of two census tables with 10,000 records and 12,000 records respectively would be calculated as follows:

- Potential matches from five algorithms (75 per cent of 10,000) 7,500
- Review time for 7,500 potential matches at 200 per hour 37.5 hours
- Likely true matches from the process 3,750

Figure 5 Example of census record linkage (false)

Kingston - Review Record Linkage

E D/Page/Line 06 0002 16 1871 29 0036 07 1881

Forename William William

Surname Brockwell BROCKWELL

Rel Head/Status Son Son Unm

Address Cambridge Pl Ham Common

Parish of Birth Kingston Kingston

County of Birth Surrey Surrey

Occupation Scholar ---

Age/YOB/Booth 6 1865 11 01 01 16 1865 01 SS_Sndx,YOB_StdC

Match Status: True False Review Initial

Sort Seq: ED Seq, Name, Match/ED, Match/Name

EXIT

1871

F2	F1	Forename	Surname	Relatio	Marit	Age	Y O B	Town Born	County	Occupation
		James	Brockwell	Head	Mar	40	1831	Kingston	Surrey	Labourer
		Lucy	Brockwell	Wife	Mar	37	1834	Kingston	Surrey	
		Frederick	Brockwell	Son		11	1860	Kingston	Surrey	Scholar
		Emily	Brockwell	Daur		9	1862	Kingston	Surrey	Scholar
		Ellen Jane	Brockwell	Daur		8	1863	Kingston	Surrey	Scholar
	01	William	Brockwell	Son		6	1865	Kingston	Surrey	Scholar

1881

F2	F1	Forename	Surname	Relatio	Marit	Age	Y O B	Town Born	County	Occupation
53257		John	BROCKWELL	Head	Mar	54	1827	Norbiton	Surrey	Blacksmith
53258		Martha	BROCKWELL	Wife	Mar	51	1830	Ham	Surrey	---
53259		James	BROCKWELL	Son	Mar	30	1851	Kingston	Surrey	Smith And Farrie
53260	01	William	BROCKWELL	Son	Unm	16	1865	Kingston	Surrey	---
53261		Jannie	BROCKWELL	Daur	Unm	7	1874	Ham	Surrey	Scholar
53262		Charles	BROCKWELL	Grandst	Unm	5	1876	Norbiton	Surrey	---

Record: 7 of 56 (Filtered)

Table 3 Illustrative processing times for Blechingley

	Number of records	Average keyboard input rates	Keyboard input hours required	Hours to check and validate	Hours to create or update standard tables	Hours to perform record linkage
Census records	11,000	60 per hour	180	40	10	40
Baptism records	2,900	100 per hour	29	6	1	20
Marriage records	500	30 per hour	17	4	1	10
Burial records	2,500	120 per hour	21	4	1	20
Local directories	730	120 per hour	6	2	0	6
TOTAL			253	56	13	96

Table 3 gives estimates of the time spent to complete all the processing for Blechingly. Whilst these can be taken as a guide, for other communities they may well vary with the complexity and diversity of the records.

Each record in a census table has two fields named 'LinkNext' and 'LinkPrev' which, on completion of the review process, are updated as appropriate with the UID of the paired record in the next or previous census respectively. When all census tables have been updated, records are then appended to a person table for the entire first census. For subsequent census tables, records are only appended if LinkPrev is empty, that is, they have no link to a previous census. For each record in the person table the average year of birth from the census records is calculated and added as a field to the table. Fields are also updated on the person table with the personID of mother, father and spouse based on the relationship to head in the census tables. When a previously identified person is missing from an intermediate census but reappears in a later census, two records will have been added to the person table. Successful links between non-successive censuses cause the merger of the two person records previously created.

There is also an *alias table*, which has fields for surname, Standard Forename, Soundex Surname Code and personID. Records are added to this table whenever a name differs from a previous census entry for the same person. Hence all records in the alias table point to an entry in the person table. At this stage, for any given name the alias table will point to an appropriate record on the person table and this feature is used for the record linkage of the other types of data. Records are linked through the Standard Forename, and Soundex Surname Code in the alias table to records in the person table, whereupon potential matches that conform to further criteria specific to the type of data are extracted for review. All household census data for the potential match are also extracted and displayed during the review process. The Access form for linking groom marriages to the person table is shown in Figure 6 and a similar layout is used for the other specific data sources. Methods of linkage for each source are now described.

Baptism records. Besides the name of the person being baptised, baptism records sometimes contain a date of birth which allows validation of the potential link. The records also contain the name of each parent together with the abode and occupation of the father, and linkage can be confirmed by comparing these with the census records displayed. When identified, the person IDs of the mother and father are added to the baptism record together with that of the individual. Even when there is no match for the individual i.e. the person baptised has never appeared in a census record, the mother and father fields may be updated if known. Additional person records are created for these non-matched individuals who may well have died or moved before the next census.

Marriage records. Each marriage record contains three possible matches: the groom, the bride before marriage, and the bride after marriage. Thus the matching process has three passes through the data. Because they are

linked through the person tables it is possible to associate individuals who do not appear in the immediately preceding census (such as girls who migrate temporarily for domestic service). Age and marital status, as shown on census records either side of the marriage date, permit a degree of validation whilst father's name and occupation can be checked against census records during the review process. At the end of the process, five fields may have been updated with a person ID, namely: Groom, Bride Pre, Bride Post, Groom's Father and Bride's Father. Where different person records then exist for Bride Pre and Bride Post, the entries are merged to a single person record.

Burial records. For burials, name and age are the only variables available for linkage. Having identified a potential match using the alias table, the person table is checked for validity. An average year of birth with a variance of more than five years from the burial record will cause that potential match to be rejected. We tend to say 'true' only when there are no competing entries and there is other corroborative evidence, such as absence from the next and subsequent censuses (especially where the previous spouse is present and is widowed or has remarried). A true match during the review process will cause the person number to be added to the burial record and the date of burial inserted in the person record.

Street/trade directories. These may contain details of occupation but are often restricted to date, name and place of abode. Record linkage may assume the person to be over 21 but has little else for validation. The review process linked via name may allow the address and occupation to be confirmed from the census records. A true match during the review process will cause the person number to be added to the record in the street/trade directories table.

Our normal approach to the review process is to present the researcher with grouped potential links, the groups relating to the same person in the first table. From our experience of census linking the use of quite strong algorithms restricts the number of potential links in any group, with very few groups having more than five records and most having just one. The weaker algorithms necessary for the linking of other sources creates much larger groupings, sometimes up to 100. In these cases the review form has an option which we call 'Auto advance'. If this option is switched on, then whenever a potential link is classified as true all other potential links in the group are automatically classified as false and the routine advances to the next group. We are currently investigating a weighting system which will put the more likely matches at the front of each group.

Creating life histories and family trees

When all linkage processing is complete, within the person table is the wherewithal to create life histories and family trees. Each census record has

Figure 6 Example of groom's marriage record linkage

MLViewGroomGrey : Form

ALL SAINTS, KINGSTON **RECORD LINKAGE OF GROOM**

John Richard ABBOTT full age Bachelor Eliza Mary WADE of age Spinster
 Secretary Kingston Kingston

Marriage Date **07 August 1876** Person ID of Potential Match **15783**

Groom's Father **Thomas ABBOTT** Person ID **China Dealer**
 Bride's Father **George WADE** Person ID **Boot Maker**

True **Match**
 False
 Review **Status**
 Initial

Auto Advance
 Current Person Code for Groom
 Current Person Code for Bride Pre Marriage
 Current Person Code for Bride Post Marriage

1861	15780	Thomas ABBOTT	35	1826	Head	Mar	China & Glass dealer
1861	15781	Ann ABBOTT	36	1825	Wife	Mar	
1861	15782	Thomas ABBOTT	10	1851	Son		Scholar
1861	> 15783<	>>John ABBOTT<<	9	1852	Son		Scholar
1861	15784	Ann BARK	76	1785	Mother i	Widow	Lodger
1861	15785	Louisa STEGATT	15	1846	Servt	Un	General Servant
====	=====	=====	====	=====	=====	=====	=====
1871	15781	Ann ABBOTT	49	1822	Head	Mar.	China & Glass Dealer
1871	> 15783<	>>John R.ABBOTT<<	19	1852	Son	Unm.	Solicitors Clerk
1871	36700	Elizabeth LUNSDON	17	1854	Serv.	Unm.	General Servant
====	=====	=====	====	=====	=====	=====	=====
1881	> 15783<	>>John R. ABBOTT<<	29	1852	Head	Mar	Secretary To Traders
1881	66772	Eliza M. ABBOTT	28	1853	Wife	Mar	—
1881	66773	John W. ABBOTT	3	1878	Son	Unm	—
1881	66774	Albert G. ABBOTT	2	1879	Son	Unm	—
1881	66775	Sydney E. ABBOTT	0	1881	Son	Unm	—
1881	66776	Elizth. BAIGENT	20	1861	Servant	Unm	Servant (Dom)
====	=====	=====	====	=====	=====	=====	=====
1891	> 15783<	>>John R.ABBOTT<<	39	1852	Head	M	Secretary
1891	66772	Eliza M ABBOTT	37	1854	Wife	M	
1891	66773	John W ABBOTT	13	1878	Son		
1891	66774	Albert G ABBOTT	12	1879	Son		
1891	66775	Sydney E ABBOTT	11	1880	Son		
1891	82488	Cecil ABBOTT	9	1882	Son		
1891	82489	Mable H ABBOTT	5	1886	Daughte		

a person ID, as do records in the tables for baptism, marriage, burial and street/trade directories. For baptisms and marriages these may also include the person ID for the parent. Once the person ID is known, then using the principle of foreign keys described earlier, all relevant records can be

retrieved for either life history reporting or further analysis. For family trees, using the person ID for father, mother and spouse contained in each person record allows the creation of generation charts as shown in Figure 1. Because both life histories and family trees are created automatically using standard features within Access there could be some minor discrepancies within parental relationships.¹⁹

In line with good database practice, life histories are not stored but are created when needed. To create a life history for a named individual, the person ID of all persons matching that name is obtained by interrogating the alias table. This gives access to the associated records in the person table whence access to all other data tables is available. Appropriate records are extracted and arranged in date sequence for printing (see Figures 2 and 3).

Conclusion

This paper has described how the Kingston Local History Project set about the preparation of basic life histories for two late-nineteenth century communities in Surrey. The use of expensive computers or bespoke computer programming was precluded by limited or non-existent funding, as was any attempt at the sophisticated processing described in several previous exercises in record linkage. The result is a set of practical procedures that can be executed at a minimal cost on any modern personal computer: there is no reason why such life histories cannot be compiled for any community in England, Wales or Scotland. No special programming skills are required although an understanding of the concepts of Microsoft Access is desirable. Reasonably accurate estimates of the time taken by the processing procedures are available, allowing budgets to be calculated and adhered to. Such projects need not be the work of academic institutions and there is no reason why groups focussed on local or family history could not set up and manage them. Indeed, through the offices of two interested local history societies in other parts of Surrey, relevant data for two more villages and a small town are currently being keyed in and will be available as life histories and family trees to their members and other interested parties.²⁰

Acknowledgements

All census data are Crown copyright and are published by kind permission of the Controller of Her Majesty's Stationery Office. The Kingston Local History Project is grateful to the staff at the Local History Room of the Royal Borough of Kingston's Museum and Heritage Service for their assistance in collecting the Kingston data and also to the excellent group of volunteers who spent many hours inputting and checking the data. Our particular thanks to Bill White for inputting all the Blechingley data and to the Blechingley Conservation and Historical Society for providing the relevant source material and checking the results.

NOTES

1. Our experience may well be standard for collaborative projects of this nature: see N. Goose, 'Participatory and collaborative research in English regional and local history: the Hertfordshire Historical Resources Project', *Archives*, **22** (1997), 98–110 for details of a project which had similar problems in its early days. More details of the Kingston Local History Project can be found in P. Tilley and C. French, 'From local history towards total history: recreating local communities in the 19th century', *Family and Community History*, **4** (2001), 139–49; and P. Tilley and C. French, 'Record linkage for nineteenth-century census returns: automatic or computer aided?', *History and Computing*, **9** (1997), 122–33.
2. J.D. Marshall, *The tyranny of the discrete: a discussion of the problems of local history in England*, (Aldershot, 1997), especially 2–3.
3. The occupational coding structure is explained in W.A. Armstrong, 'The use of information about occupation', in E.A. Wrigley ed., *Nineteenth-century society: essays in the use of quantitative methods for the study of social data*, (Cambridge, 1972), 215–23.
4. See E.A. Wrigley, R.S. Davies, J.E. Oepfen and R.S. Schofield, *English population history from family reconstitution 1580–1837*, (Cambridge, 1997).
5. B. Eckstein and A. Hinde, 'Measuring fertility within marriage between 1841 and 1891 using parish registers and the census enumerators' books', *Local Population Studies*, **64** (2000), 38–53.
6. See D. Courgeau and E. Lelièvre, *Event history analysis in demography*, (Oxford, 1992). G. Alter and M. Gutmann, 'Casting spells: database concepts in event-history analysis', *Historical Methods*, **32** (1999), 165–76, describe a recommended approach for organizing the data to be used in such analysis.
7. There are also markers to indicate the last record of a household and the last record of a dwelling house.
8. For a guide to trade directories, see D.R. Mills, *Rural community history from trade directories* (a *Local Population Studies* supplement), (Aldenharn, 2001).
9. With a large proportion of the data being on microfilm or microfiche, printing charges were kept to a minimum by the generosity of the Royal Borough of Kingston's Heritage Centre who allowed us to take hard copy, the only cost being paper.
10. Armstrong, 'The use of information about occupation', 215–23.
11. A detailed explanation can be found in I. Winchester, 'What every historian needs to know about record linkage for the microcomputer era', *Historical Methods*, **25** (1992), 149–65.
12. See, for example, the papers in E.A. Wrigley, ed., *Identifying people in the past*, (London, 1973).
13. J.E. Vetter, J.R. Gonzalez and M.P. Gutmann, 'Computer assisted record linkage using a relational database system', *History and Computing*, **4** (1992), 34–51; C. Pouyez, R. Roy and F. Martin 'The linkage of census name data: problems and procedures', *Journal of Interdisciplinary History*, **14** (1983), 129–52; P. Adman, S. Baskerville and K. Beedham, 'Computer-assisted record linkage: or how best to optimize links without generating errors', *History and Computing*, **4** (1992), 2–15; H.R. Davies, 'Automated record linkage of census enumerators' books and registration data: obstacles, challenges and solutions', *History and Computing*, **4** (1992), 16–26.
14. Davies, 'Automated record linkage'. Davies's procedures were implemented on a mainframe computer using the package Scientific Information retrieval (SIR).
15. C. Harvey, E. Green and P. Corfield, 'Record linkage theory and practice: an experiment in the application of multiple pass linkage algorithms', *History and Computing*, **8** (1996), 78–89.
16. Tilley and French, 'Record linkage', 131.
17. *Short Standard Forename* is based on the first three letters of the Standard Forename and is held in the standard forename lookup table. Where desirable it is adjusted: hence Bill, Will, Wm and William are recorded as 'Wil'; whilst Hannah, Ann, Anna, Anne and Annie are all recorded as 'Ann'. The *Soundex Surname Code* is a phonetically based code designed to pick up matches of similar sounding names. It is considered by some to be inappropriate to nineteenth-century handwritten records. We do not accept this as many of the names are written as dictated. Some 15 per cent of our linked records match on Soundex but not on full surname. It is, however, crude, and whilst it will match Smith with Smyth it will also match Smethurst with Saunders. The Soundex code is implemented within Access queries by using a pre-recorded programming

routine called a function.

18. The *Guth matching score* is a measure of compatibility between two names and is derived from a pattern matching technique designed by Gloria Guth in the 1970s: see G. Guth, 'Surname spellings and computerised record linkage', *Historical Methods Newsletter*, **10** (1976), 10–19. Since the names are not coded in any way it can generate an enormous number of potential matches so records should be filtered before the technique is applied. It is best used on residual records where other algorithms have failed.
19. Children of lodgers are often erroneously shown in the CEBs as children of the head of the household. Other cases show children as sons and daughters of the head of household when both the children and their real parents are living within the household of the children's grandparents. With careful analysis, these can often be detected and the standardized relationship adjusted without changing the original data.
20. Should any readers of this journal wish to set up similar operations the author would be very pleased to offer advice and copies of standard Access routines. His email address is petertilley@talk21.com.