Priorities for record linkage: a theoretical and practical checklist

Winchester, Ian

Ian Winchester

Priorities for Record Linkage: A Theoretical and Practical Checklist

Historical record linkage, as S. Langholm has recently noted, is „a whole little science if its own". By this term we generally mean the bringing together of historical records relating to the same historical unit of interest – usually a person, but often as well, a family, a household or perhaps a process, event or object. And usually when we do this we employ routinely generated records such as parish registers.

Although this little science (or sub-science) is a new one, it is logically connected with a very much older historical practice, that of saying more than one thing about historical individuals (units) by means of reference to more than one document or record relating to that historical individual or unit. Although the logical part of this little science is shared by all historical activity (the part connected with the reidentification and further characterization of historical individuals), the technical parts are of greatest interest to the field of micro-history and its blood-brother, micro-demography. In both of these fields we are primarily concerned with building up collective biographies of individual people, or of individual families, or of individual households prior to our analysis of the resulting data. As a rule, the main sources employed for such purposes have been parish registers (France, England, Sweden, Denmark) or, for the 19th century especially, census rolls (U.S.A., Canada, Britain) and their near neighbors taxation lists or assessment rolls.

A more recent practice has come to be the use of either of these more or less total record sources as a backbone for the research in question, with the addition of as many other sources as well. In principle, there is no limit to how many record sources one might use.

Since interest in the micro-historical and micro-demographic fields is both high and increasing, and since many more individuals and groups are considering using record linkage practices in their work, I want to discuss problems of priority. That is, I want to talk about those general problems which anybody engaging in micro-historical or demographic work may face in a record-linkage context. The specific problems, while they have not been well reported, have at least been reported. And this work is readily available to the would-be record linker. In this regard I mention a book edited by E. A. Wrigley called *Identifying People in the Past*, an article by Theodore Hershberg in the *Historical Methods Newsletter* in 1976 and an earlier article by myself in the first issue of the *Journal of Interdisciplinary History*[1]. I shall

[1] Wrigley, E. A. (Ed.), Identifying People in the Past, London 1973; Hershberg, T., et al., Re-

assume that if you are really interested in this area you will read these few things. What the literature really lacks is a discussion of the variety of factors which face a researcher or research team who wish to know whether or not they should automate, partly automate or do the entire job by hand. And, furthermore, whether or not they should perform a sample linkage first, whether or not they should establish an iron-hard set of linkage rules in advance, and what general problems they should consider and possibly tackle with some zest.

1. Factors Affecting Record Linkage Priorities

The would-be micro-historian or demographer will find his record-linkage planning constrained by three obvious, but usually not well thought out, features of his task. First, by the resources that are available to actually bring about the necessary record-linkage. And, secondly, by the problems posed by the files with which he must work. In Figure 1 I have listed these two headings and their main sub-headings.

Figure 1: Factors Affecting Record -Linkage Priorities

Available resources	Staff available (pay/no pay) Data processing (free/charge) Money (quantity/years avail./strings)
Problems posed by files to be linked	Desired organization of final file Number of files to be linked Size of files Condition of files Kind of linkage (total/sample) Available identifying items (common/overlap/independent) Biases Legal or moral constraints

Each of the above factors and their sub-factors are interrelated. And each potentially affects the choice of record-linkage priorities. The basic outcomes of any such consideration of record-linkage priorities are decisions as regards a) file organization, b) file preparation, c) the linkage steps themselves. As regards the first of these,

cord Linkage, in: HMN, 9 (1976), pp. 137–163; Winchester, Ian, The Linkage of Historical Records by Man and Computer: Techniques and Problems, in: The Journal of Interdisciplinary History, 1 (1970), pp. 108–124.

we have two basic decisions to make: How to organize the files prior to the linkage; and how to organize the master file after the linkage step or steps. File preparation refers to how we wish to prepare the files in advance of the linkage step(s) – for example, by means of a number of data transformations to enable detailed comparisons of records to be made more easily. The major decisions to be made with respect to the record linkage steps are, first, whether to do the entire process by hand, or to enlist the aid of some data processing device(s) for part or all of the actual linkage, and second, what exact linkage rules and steps to employ.

Before going on to say a few things about the items in Figure 1 and their relation to the three basic kinds of decisions, I shall give, in Figure 2, a chart summarizing the latter.

Figure 2: Possible Outcomes of Record-Linkage Priority Considerations

Decision as regards		
a) file organization	e. g.	sort n-files alphabetically
		organize output files by individual
		family
		household
		parish
b) file preparation	e. g.	delete no information
		delete all nominal information
		recode all surnames
		recode all given names
		classify all occupations
c) linkage steps	1. determine exact linkage rules	
	2. choose among	fully automated
		partially automated linkage
		hand

In terms of these various factors and outcomes which affect record-linkage choices, there are two orders of interest: an order of importance, and an order of temporal priority. For example, it is a mistake to determine exact linkage rules prior to the examination of the files under consideration. And it might be a mistake to decide prior to such a consideration that one will opt for a fully hand-linked operation. So, simply because the linkage steps might be a paramount, or a high priority consideration, it does not follow that they should be determined first.

Probably the most important consideration of all is what the historian/demographer/geneticist/etc. intends to do with the linked files. Usually the answer is clear. The researcher wants to create cross-tabulations of certain kinds which form the basis of his descriptions or explanations, and which may suggest other forms of analysis or other problems. If this is so, then the first matter to which consideration should be given is the matter of the desired logical organization of the final or out-

put file. This could be in the form of a more orderly collating of separate cards full of information or of their card-image equivalents. But if the studies to be undertaken involve detailed examination of, say, married siblings or of multi-generational family groupings, then detailed consideration will have to be given at an early stage to the graphical or list representation of the output file.

Perhaps the second most important consideration, next to that of knowing roughly what you want to do with the linked files, is that of getting acquainted with the files in a detailed way. It is really crucial at an early stage to get to know the quirks, the strengths and weaknesses and biases, of one's files. There are exceptions to this rule. But it is really very important for the would-be record linker to plan a honeymoon with each of his files in turn prior to linking them, rather than to couple them together into a harem and then imagine that detailed familiarity will come more easily later. None of the record linkage decisions listed in Figure 2 are possible unless a researcher is acquainted in a technical way with his files prior to any linkage on a big scale. The sort of technical way I have in mind is detailed in the articles and book I mentioned earlier. But in summary they are these:

What identifying or potentially identifying items are common to or overlapping on the files to be linked?

With what relative frequencies do the items occur? (e. g. Are there more Browns than Schmidts and more Russells than Johannssons?)

What varieties of names are there with similar spellings, translations (White-Blanc-Bianco), or transformations because of dialect, patois or identifying necessity?

With what frequencies are identifying items likely to be discrepant on linked pairs of records?

Exactly how large are the files? What is the distribution of the various identifying item frequencies in each item kind? What is the range of such frequencies?

If one has this kind of information about one's files at an early stage, thinking about file preparation, organization and the detailed linkage steps is made much easier. This ease is partly because of the mere fact of detailed familiarity. But it is also because each of these other matters requires decisions based on the kinds of knowledge referred to above. Most of us who have set up rather large data bases involving a linkage component have stumbled into the matter without such prior detailed considerations of our files. We did not know that it was needed, nor could we imagine before the fact why such detailed consideration would be needed. If one has such knowledge at an early stage, then one can use that knowledge for the following purpose:

To estimate the time required to complete a particular portion of the linkage job.

To estimate the cost of completing a linkage job by hand or by machine or partly by both means.

To estimate biases in one's data or in potential sampling procedures or in potential linkage procedures.

To determine how to prepare the files as regards standardization, transformation of identifying items such as names into a standard format for sorting or detailed comparison purposes.

To estimate the relative weights to be placed on the agreement or disagreement of various identifying items during a linkage step.

And therefore:

To determine, or help determine, the exact linkage rules needed.

And finally, because of all these things:

To help the researcher choose among a fully automated, partially automated or completely unautomated linkage procedure.

The third most important consideration is that of knowing as exactly as one can (or being as scrupulously honest as one can) about the matter of available resources for linkage – including file preparation, detailed planning, programming and supervising in terms of money, people or free computer time or free programming help. This assessment is an enormously important consideration. But it ranks third in my suggested list because I am an optimist who in his heart believes that where there is a will a way can be found. It also ranks third because, if you do not have any idea about the problems you want to tackle first and if you are not acquainted in detail with the files you want to use in the manner I have suggested, then you cannot begin to know if the resources you have available are enough to enable you to do the job. There is room for a little circularity here, of course. For if you do not know that you have some resources, then you cannot plan even to have the detailed familiarity with your files that I suggest you have. Here, however, I think that the historian's traditional resource (himself and perhaps a graduate student or two) will probably suffice in a pinch.

I cannot really hope to give much useful advice here on the matter of comparing needs to the resources available. So I will offer, instead of something theoretically and practically satisfactory, a comparison of a number of successful projects and of the resources which they have available relative to the tasks they are pursuing.

I have chosen five projects with their record-linkage tasks. The five I have chosen cover a fair range of the spectrum of tasks and problems, as well as a fair range of file sizes and problems. In the order of their reporting of their record linkage techniques and problems they are: The Hamilton Project, The Cambridge Group, the Philadelphia Social History Project, the Umea Demographic Database, and, in order to illustrate small files, Stewart Hardy's Model School Study.

The Hamilton Project has reported some of its methods and problems[2]. The Cambridge Group have published a long discussion in the book edited by Wrigley mentioned earlier[3]. The Philadelphia Social History Project has recently published an article which builds mainly on Winchester's 1970 article, but contains a very interesting discussion of bias using census to census linkages[4]. The Umea Demographic Database has reported on its files in detail. There is no discussion of record-linkage as such in these reports, because, I believe, there is a belief that there are no

[2] Winchester, op. cit.

[3] Wrigley, E. A., and Schofield, R. S., Nominal Record-Linkage by Computer and the Logic of Family Reconstitution, in: Wrigley, People, pp. 64–101.

[4] Hershberg et al., Record-Linkage.

problems[5]. If true, this is a remarkable fact. Finally, Hardy is a doctoral candidate who has nearly completed his work on Model Schools in Ontario in the last century. His work exemplifies small, multiple file linkages[6].

What we want specially to dwell on is the match between the resources available and the magnitude of the task faced by each researcher or team.

Both the Cambridge Group and the Demographic Database work with parish records. The Hamilton Project and the Philadelphia Social History Project work with 19th century census rolls and other parallel sources. If I were to rank them in order of size in terms of the sheer data which the projects handle the order would be:

Demographic Database	Gigantic	Approx. 10^7 records
Philadelphia Project	Very Large	10^6
Cambridge Group	Large	$5x\ 10^6$
Hamilton Project	Medium	10^5
Model School Project	Small	$3x\ 10^3$

where by a ‚record' I mean an „80-column card image".

In terms of the difficulties which the data pose, however, I would give quite a different ranking, rather more like:

Cambridge Group	Extremely difficult	only names, many variations, bad handwriting
Hamilton Project	Very difficult	25 % of items with discrepancy on matched pairs, handwriting difficult
Philadelphia Project	Very difficult	similar to above
Model School Project	Fairly difficult	only names and ages (sometimes)
Demographic Database	Fairly easy	often whole families, much linkage done by priest at time

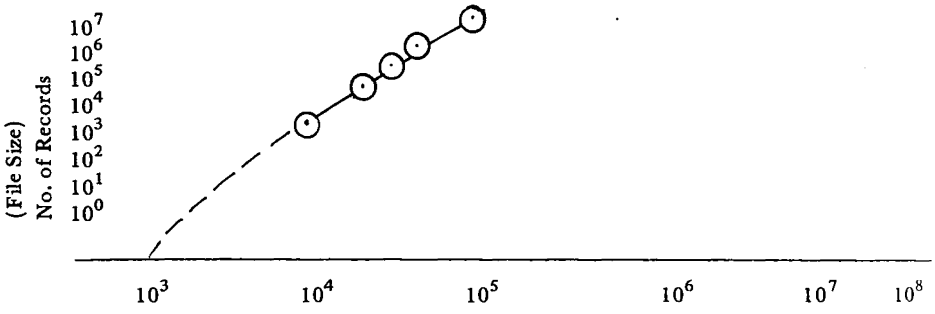
In terms of the resources available, since these are projects in progress or completed, there is a direct correlation between the resources available and the size of the project files. My best guess as to the ranking, personnel, money available and computing power is a follows:

Database	*Personnel*	*5-year funds*	*Computer power*
Umea DD	Clerical 40 Technical 1 Supervis. 2 Historian 1	$\$ 1.5 \times 10^6$	Tape to disc IBM 360 No extra cost

[5] Johansson, Egil, and Sundin, Jan, The Demographic Database, I and II, mimeo. Umea 1977.
[6] Hardy, Stewart, Linking Educational Records to a Manuscript Census, mimeo. Ontario 1977.

Philadel.	Clerical 4–12 (part-time) Technical 2 Sup./Hist. 1	$ 5 x 10^5	Card input Interactive setup IBM 360 Extra cost
Cambridge	Clerical 2 Technical 1 1/2 Historical 3	$ 2.5 x 10^5	Paper-tape input Nottingham IBM 360? Extra cost
Hamilton	Clerical 1–2 Technical 1/2 Historian	$ 1.25 x 10^5	Card input IBM 7094/360 No extra cost
Model Sch.	Historian 1	$ 2.0 x 10^4	IBM 360 & Minicomputer No extra cost Hand calculator

If we were to graphically display the relationship between the funds and the file size it would look something like this:



Cost in Dollars for 5 years

Of course, the costs I have graphically displayed in comparison with the file sizes are the total costs of running a viable historical research project. They are not just the costs of recording the data in a usable form (except for the Umea project). It would be possible to project costs going the other way – namely from unit costs estimates and size of job to costs for a project of a particular size for a five year period. But I think that these rough approximations to reality suggest quite clearly how expensive a large database using linkage techniques is.

It is probably unfair to try to do a ranking of productivity in terms of scholarly product of the various projects listed. But it would be reasonable to expect that projects with a higher number of historical-academic staff would be more productive in a given five year period. In this regard, I will invent as a publishing unit a „Russell" which is the rate at which Bertrand Russell published in a year . . . roughly one book and ten articles, for a total of perhaps 500 pages of print. At this rate, the Cambridge project has been publishing at roughly a Russell/year or more, for a total

of possibly six or seven Russells in a five year period. The Hamilton Project is pretty close behind in the period 1970–1975 with five annual reports of some 300 pages each, a number of independently printed articles by Katz and Winchester and a book by Katz for possibly a 4.5 to 5 Russell total. The Philadelphia Project in the period 1972–1977 is loping along at about half a Russell a year, but is picking up the pace rapidly in the last year or so for a five year total of some 3 Russells. However, recently a flood of scholarly workers is on the scene. The Umea project has put almost all its effort into the data transcription, file preparation and linkage process in the past five years. Egil Johansson's work on literacy has been the main research using the database to date. And while this seems to me to be the most significant, perhaps, of all the work done, it does not amount to more than a Russell or two all in all. Hardy's work has so far issued in a couple of papers[7].

The point of this aside is absence of necessary correlation between effort in studies involving record linkage and either quantity or quality of results. Consequently, while it is certainly necessary to see that the total resources one has available are sufficient for the task at hand, I think it is still of higher priority to know what one wants to accomplish with the research in question and to be acquainted with one's sources in intimate detail. Then the resources can be found, or often can be. Or else one can tailor one's research to the resources one has. But to do this requires that the resources and the technology be made a third-rank priority.

So in summary, I would list the priorities for the record-linker to be as follows in rank order:

1. Knowledge of what is to be accomplished, of the kinds of problems to be tackled;
2. Intimate knowledge of one's source files;
3. An honest appraisal of one's five year resources.

I am strongly recommending, therefore, that the concepts and problems come first and the technology a distant third.

2. To Automate or Not to Automate, That is the Question

Having given my global priority recommendations I will now try to say some words about the vexing problem of when to commit oneself to help by a machine and when not do so. Three rules of thumb suggest themselves to help one decide whether to link by hand or to attempt to automate the linkage. These rules are as follows:

1. If the files are very messy and the identifying items few, then one should link by hand.
2. If the files are very large, and the identifying items more than adequate, then one should link by computer or partially so.
3. If the files are small, then one should always link by hand.

[7] Hardy, Stewart, Educational Records.

As general rules of thumb (as opposed to general rules) these three are pretty good. But they are certainly inadequate to cover all cases. I shall attempt to dicsuss those cases in which they are inadequate, because it is here that mistakes can be costly of time and effort.

In the case of the first rule, a great deal hinges on the notions ‚very messy' and ‚few identifying items' which are not, as they stand, very scientific. Even if there are 20—25 % errors in the identifying items which one has, there exist perfectly good methods for reducing the potential effect of these so that one can by automated means achieve roughly the same result as by an intelligent filing clerk. I shall not go into details here since the matter is pretty well covered in the literature[8]. The basic idea here is simple. If, for example, names relating to the same individual have a certain tendency on the average to be differently spelled, one can counteract that tendency by systematically transforming the names on the files to be matched to a kind of common spelling. (For example, Smith, Smyth and Smythe might be transformed to SMTH.) Similarly, if other potentially identifying items (such as age, for example) are often in disagreement when two records relate to the same person, and if we can estimate the frequency and the extent of disagreement, then we can devise data transformation schemes to compensate. This method, of course, presupposes a familiarity with one's files of the sort which I mentioned before — very likely in the form of a small hand-linkage study of them. Suppose, for example, that you find that ages can often be discrepant by as much as five years on pairs of records that relate to the same person. And furthermore, suppose that it is not uncommon that the figures in the ages are transposed in the inscription: ‚25' and ‚52'. Then we need, at the time we come to compare identifying items on a pair of records we are considering for linkage, two transformation rules. One is that of considering the age on one record and comparing it with that on the other in such a way that we consider it evidence for the linkage if the ages are within five years of one another. The other is that of comparing the ages and their digit reversals. If one is the reversal if the other, then we consider this evidence in favour of a link. If the frequencies of either of these occurrences are known or can be estimated from a sample linkage, then we can weigh the evidence[9].

I would, myself, follow 1. and do the linkage by hand if I thought that the files were so messy that no sample of them would enable one to predict what was coming in the next batch. And I would similarly do the linkage by hand if my computing and clerical resources were minimal. Otherwise, I think one should do as much of the work by mechanical or electromechanical means as one can.

[8] See, e. g., Winchester, Historical Records and Hershberg, Record-Linkage.
[9] See Winchester, op. cit.

2.2. *Large Files*

It seems obvious that if the size of files is very large and the identifying items more than adequate, then one should link by computer. I still think this a pretty good rule of thumb. But I no longer think that it is a general rule that one abandons at one's peril. There are two factors which can throw this obvious rule into doubt. The first is lack of computing and programming power. And the second is excess of clerical power. If you do not have a computer available, preferably free, with adequate storage facilities, large enough core capacity, then you would probably do better to take a decade and do the job by hand, writing articles as you go. This method is certainly better than trying to manage on an inadequate budget with an inadequate machine. Another ground for doubting the wisdom of this advice is if you have foreknowledge that in a year or two the machine you have (and the programmers or systems people) is going to be obsolete, or transformed into a minicomputer, or sold. It seems to me better to take five years out of your life and have you and your graduate students do the job by hand than to constantly have to reprogram or become acquainted with new machines, people and their quirks. This situation is a general problem recognized in the computer industry under the general heading „the interface problem". And it is a genuine problem and a very great bore for historians – though, of course, a lot of fun for computing people. Of course, if you have the money, the computer and adequate programming backup, then by all means automate or semi-automate.

If you have an excess of clerical staff, then you may do well to let others do the automating. This course of action requires, for very large files, that a number of other conditions be met. The only place where I know that this has successfully been done is at the Demographic Database in Umea, Sweden. The files being used in this case are beautifully kept parish records from the 19th century which are systematically linked to other registers maintained by the clergy from the 1740's following a Royal Decree of 1748. Essentially a special register of the capacity of each household to read and write – by means of graded marks – is linked to vital statistics registers. Each time an individual appears in one of the five basic sources an excerpter fills in a card. At a later stage in the operation all of the cards relating to a single individual are sorted together by hand before being keyed to a disc storage device for later data processing by computer. Since the system is based on individuals, a so-called „guide-card" is also produced by hand which gives, for every individual, his own identify, his father's, his mother's and the identify of the husband and wife. Links are thus made from son to father and between husband and wife, thus facilitating simple family reconstructions at a later stage.

Since there are forty clerical staff involved in this process, and since there is little difficulty in making the links for anywhere except Stockholm or between Stockholm and another parish, the linkage work can be done very quickly and at the same time as the original excerpts are made. Thus even though there are excellent computer facilities available in Umea, I see no reason why the linkage portion of this mammoth operation should be automated. Though, perhaps, when Stockholm is tackled some time in the next century one might want a little machine assistance!

This commonsense rule of thumb runs „If the files are small, then link by hand". And for two small files (say two voters lists in Cambridge in successive election years in the mid 1800's) this is certainly a good rule. But if more than two files are under consideration, the number of detailed comparisons between record pairs which one may have to face (especially in parish register work with only nominal identifying items) can become astronomical. The difficulty which one faces by hand-linking files which are essentially unordered is that one has, in effect, potentially to compare every record to every record in each of the other files. Even with only three distinct files with five records in each file the number of possible linkages is large, with the number of distinct persons possible ranging from five to fifteen. If (in the worse possible case) all of the names were similar to one another and one had a limited amount of information to enable the resolution of ambiguities, one would have to consider 12,962,661 possible linkage arrangements[10]. In the best of all possible worlds, one would still have to consider 125 possible links. And as the number of files increases, the number of links possible goes up exponentially.

Clearly, even in the three file case, something must be done – to bring the number of comparisons down to feasible proportions. The most obvious and most used strategy is to reorder the data in each file alphabetically or numerically and to limit the detailed comparisons between records (among records) to those within a limited range or „pocket" within the sorting key. For example, with linkage mainly dependent upon surnames and forenames one might sort all files by surname and only consider for detailed consideration those surnames in each file which have the same surname. For very messy files, especially small ones, this can be a disastrous manœuvre since one might lose twenty percent of one's actual links. Thus, again, one might be driven to some sort of coding of surnames which will bring together all of those surnames which are sufficiently similar to warrant a detailed comparison. If this sort of data transformation is too strict, it misses most of the links. But if it is too loose, one is back to comparing nearly all records with all records. Here there is room for art! Most of this sort of thing can be managed with only a card sorter and a keypunch machine for small files. But for many of these, a computer is a very useful aid. So while the rule of thumb is pretty good, it has its notable exceptions.

One final note under this general heading. If the files are very messy, the identifying items few and their quantities very large, then one would be wise to abandon the task. But since I am describing the situation facing those using parish registers in England and in many other countries, or censuses prior to 1840 pretty well everywhere, perhaps my best advice would be to befriend an oil sheik who has a passion for microhistory and demography.

[10] Skolnick, Mark, The Resolution of Ambiguities in Record-Linkage, in: Wrigley, People, pp. 102–127.

3. Definite Linkage Rules for All Files

I would now like to offer a rule of thumb which I think should be given the dignity of a general rule in present day historical practice. *That is, all linkages, whether small or large, with messy or clean files, with computer or by hand or partially automated, should follow definite rules which are reported as a standard part of the project.* This is a rather long and somewhat priggish rule of thumb. It is a recommendation for which I have propagandized in support of both with my students and with anybody else who sought my advice on their record linkage problems.

My prime reason for this bit of stubbornness is this. It is only if one knows in advance the linkage rules which one is applying, that one can consistently link records such that one's linkage can be reproduced by another scholar. Now, just as in chemistry one should rightfully specify one's methods and processes so that they can be reproduced by someone else, so should one do the same thing in history. Such a simple step as this would go a long way towards increasing the respect for microdemographic or microhistorical research among other scholars. Certainly the beauty of one's descriptive prose need not be affected by such a step. And the plausibility of one's argument might be increased considerably.

Furthermore, although such a simple reporting of one's record linkage rules would enable another scholar or scholars to reproduce and check one's work, the result would likely be quite the reverse. Since one could check another's work at any time, one need not. Trust begins in shared methods. Some of the most strident controversies of the last number of years involving 19th century American cities and their social structure have arisen partly because of an inadequate reporting (as well as thinking through) of methods.

There are also reasons for having such definite linkage rules formulated and followed, even for the benefit of the scholar who is himself using them. The most obvious reason is the ease and comfort which systematization brings in its train – basically a gain in confidence and clearheadedness. But the reason of more scholarly, rather than psychological, import is that it will enable the estimation of systematic biases.

Of course historical sources of the routinely generated kind are always biased. Ecclesiastical parish registers are denominational. Thus non-Anglicans in England, non-Catholics in France will either be missing or underrepresented in parish registers. The only exception I am aware of is that of an established church which tolerates, for statistical purposes, no exceptions – such as the Lutheran Church in Sweden between 1640 or so and 1900. Tax registers tend to ignore the poor or to cover them less fully. Differences of sex, marital status, age and the length of residence, or quality of health, may all lead to differential coverage in the kinds of records which we may wish to link.

But record linkage processes, as such, can compound such biases or create new ones. The crucial issue for our discussion here is what kinds of biases record linkage processes do or may add to the initial circumstances. Now as a rule, and excluding

the Swedish work alluded to above, any record linkage process can be seen as a sampling process in practice.

This process can be seen by the following illustration. Suppose we wish to compare the incomes with the family sizes of the various scholars who engage in record linkage practices. (Perhaps we have a theory about the effect of such practices both on income and on fertility!) Our first file looks like this:

File A *Incomes of Record Linkers*

Surname	Given Names	Income (Annual)
Henry	Louis	250,000 fr
Wrigley	E. Antony	5,000
Schofield	Roger	4,500
Newcombe	Howard B.	30,000
Felligi	Ivan	47,000
Winchester	Ian	25,000
Hershberg	Theodore H.	100,000

and the second like this:

File B *Family Size of Record Linkers*

Henry	Louis	27
Sunter	Ian	2
Kennedy	James M.	7
Skofeld	Roger	2
de Wyncestre	Iain	1
Johansson	Egil	3

I think that, using all our wiles, we might come up with the following result for these two files, namely:

Linked File: Incomes and Family Size of Record Linkers

Henry (Henri)	Louis	250,000	27 children etc.
Schofield (Skofeld)	Roger	4,500	2
Winchester (de Wyncestre)	Ian (Iain)	25,000	1

Our problem as historians using record-linkage techniques is how to use this sample, generated by the vagaries of the record-linkage process, as representative of the entire lot of record linkers. If the sample is random, we can do quite a lot. But if there are systematic biases in our files of which we are not aware, then we stand the risk of producing a systematically biased, non-random sample — one which we have no clear way of handling.

Furthermore, if we are dealing with very large populations, then we may wish to sample the files to be linked in advance. The net result of the record linkage process could be the sampling of a sample. And depending upon the initial sample chosen it could be a biased sample of a biased sample. I do not know of any discussions in

the statistical literature which exactly parallel this process. Felligi and Sunter[11] suggest ways of handling intercorrelations among identifying items which may destroy the randomness of the linkage process, central to their mathematical model of the process. But what we are talking about are biases due to the historical, rather than the statistical, qualities of the data.

The problems can best be illustrated by two of the tasks which have been central to North American Research in the last decade. The first is the study of „persistence" or its inverse „population turnorver". And the second is the study of occupational, geographical amd economic mobility.

In persistence studies, the proportion of a given population persisting over time has been tied to a variety of significant historical features. Persistence can function both as something to be explained and as something which, if given, explains other features in the data. Thus, if rates of persistence are known for various communities, we can inquire how these are affected by differences among such communities, such as size, location, history, age, economics, rates of growth, population composition, access to cheap transportation and the like. As an independent variable, degree of population turnover has been used to explain all of the following[12]: lack of class-consciousness, limited working-class militancy, the slow growth of labor organizations, community instability, a variety of social pathologies, the continuing control by a small social elite and a stable social structure. There are two problems connected with record-linkage processes here. The first is that for any such persistence study we need to be reasonably sure that our rates of persistence are neither gross overestimates nor gross underestimates. And estimated rates are a function of the nature and manner of the recording of the data, the quality of the data for purposes of linkage and the exact record-linkage algorithm used.

The second problem is whether or not the qualities of those who persist differ significantly from those who do not. If they do not, then it is the mere fact of persistence in the community which is under study. But if they do differ, then it is reasonable to inquire into the special causal circumstances which may obtain. Again, in judging such research, both one's own and that of others, the quality of the reporting of the record-linkage processes is crucial.

In social mobility studies another feature connected with the record-linkage algorithm used is important. Whereas in persistence studies we want to use all the variables we have available which can potentially function as identifying items, in mobility studies such use can be a source of systematic bias. Suppose, for example, one wishes to follow occupational change through time. If occupation is also used as an identifying item – and it is a very good one – then we might potentially reidentify a higher proportion of people whose occupation did not change or which changed marginally during the time period under consideration. The result would be a biased sample produced by our process. Since such a bias is hard to avoid in

[11] Felligi, I. P., and Sunter, A. B., A Theory for Record-Linkage, in: Journal of the American Statistical Association, 64 (1969), pp. 1183–1210.
[12] Hershberg, Record-Linkage.

hand linkage studies, this is another reason why standard linkage practices and reporting of these would be a great boon to workers in the field.

One final example of potential bias produced by linkage processes: In order to make linkage easier a number of researchers have used a strategy of systematically deleting commonplace names from their files. This practice certainly does make re-identification much easier and considerably reduces the files under consideration. But it also means that if it is the common man (Schmidt, Smith, Jones, Andersson and Lefebvre) who is the object of study, then the most common of all is systematically left out. There has been some discussion of Thernstrom's use of this method[13]. Standardized reporting of linkage rules and methods would have at least facilitated or, more likely, made such discussions unnecessary.

It is, I think, clear that definite linkage rules for all linkages – hand, machine-assisted and computer-automated – are a boon to our art. And I think that it is also clear that we require some standard means of reporting on our files and our rules which we all can understand.

4. Standard Reporting for Everyman

I shall not argue further here for a standard reporting of one's files and linkage rules for all studies involving record linkages in a historical context. What I want to do is simply suggest a minimum list of things which should be reported. To do this I shall refer back to Figures 1 and 2. I think that a footnote or an appendix to published work involving record linkage should mention at least the following:

A. Original Files:
   1. The number and type of files linked.
   2. The size of each file.
   3. The organization of each file in its original form.
   4. The number and kind of available identifying items.
      If there are many files, then whether the identifying items are common to all or only by file pairs.
   5. The condition of the identifying and descriptive items on each file. Emphasis should be placed on such things as surname variations and the likelihood of discrepancies in identifying items in truly linked record pairs.
   6. Systematic biases in original files.

[13] Alcorn, R. S., and Knights, Peter R., Most Common Bostonians: A Critique of Stephen Thernstrom's The Other Bostonians, 1880–1970, in: HMN, 8 (1975), pp. 98–114; and Thernstrom, Stephen, Rejoinder to Alcorn and Knights, in: op. cit., p. 117.

B. Files as organized for linkage:
   1. The preliminary preparation of each file based on 5 above. Whether or not recoding of, say, surnames or occupations has been done. Whether a derived sorting key was added by hand or by a machine step. Whether surnames were given a standardized spelling prior to linkage steps.
   2. The file organization prior to linkage. Whether this is the same as on the raw files (3 above). Whether files have been sorted according to some sorting key or other (e. g., surname, sound, by occupation code). Whether each file is organized by individual, family, household, parish and the like.
   3. Whether the physical order of the files is the same as the logical order.

C. Linkage Steps:
   1. Whether the linkage is a hand operation, a partially automated one or a fully automated one.
   2. The results of a preliminary hand linkage, if undertaken.
   3. The exact linkage rules followed.
   4. The final file organization for the linked file.
   5. The biases which the particular linkage steps may involve.

If such a reporting became a standard procedure among micro-historians and micro-demographers, both the quality and the comparability of what we do would be improved significantly. I think we would also be more convincing.

It is time to summarize what I have said, both to suggest a number of unsolved and untackled problems and to mention some future possibilities which recent technology tends to make possible. I have tried, by way of filling a gap in the literature, to discuss a number of factors affecting record linkage priorities and decisions in the context of some work involving the techniques in the last decade. I have suggested that there are three overarching priorities for any would-be record-linker: namely, that he know exactly what his problems are and why he needs such a linkage; that he have an intimate knowledge of his source files before pushing ahead to plan the linkage steps; and that he be as honest and knowledgeable about his exact requirements and resources as he can be. If these three priorities are seen to, the researcher is pretty certain to prevail. I have tried to give some examples in detail as to what the range of possibilities might be in each case.

As regards whether or not to automate, I have suggested three rules of thumb, each of which has important exceptions. These were, first, that if the files are very messy and the identifying items few, then one should probably link by hand. But if the files are also very large and one has access to clerical and computer assistance, then techniques exist which can overcome the technical difficulties. If the files are very large, the second rule of thumb is that one should think of machine assistance from the beginning. But, if one has a large clerical operation available and the linkage step is a natural one, then one can avoid using a machine – especially if the files are clean and there is hardly ever doubt about a linkage. The third rule of thumb is that if the files are small, then one should always link by hand; but if one has many such

files, the potential number of searches one would have to do increases hypergeometrically. Thus in some cases machine assistance would be helpful.

As regards the matter of definite linkage rules, even for hand linkages I have argued that one should always have a precise set of linkage rules whether or not one is using a machine for any part of the operation. We are sloppy and easily deluded animals — even in our spiritual parts. Clio is a quiet and careful muse. Reporting one's record linkage procedures in a standard fashion would be a boon to our art. I have given a brief checklist of matters which such a reporting should, at a minimum, include.

For those interested in „the whole little science" as a problem in itself, I have mentioned in passing a number of problems which we would do well to dwell upon in a technical fashion. We need a good discussion of biases and of sampling , which are creatures of the record linkage processes themselves. We need a much fuller discussion of the record linkage processes in terms of graph theory, especially since recent developments in computer science are tending to see graphical descriptions as important tools for describing many processes and files. We need somebody to draw all of what we know together into a handbook for historians and demographers, since what we have is scattered and incomplete. We need a series of standard programs for those using parish records and for those using other file types which we can use as easily as we use SPSS and Data-Text. We particularly need some interactive programs so that we can automate that which we can and can stop to look at specially difficult problems as the process proceeds.

Are there any interesting technical developments which might aid us in the future? Well, there is at least some hope that the central processors of the future will have much more storage to play with and will function at much higher speeds. But the most interesting developments, to my mind, will be in the training of young historians who will take all of what we are presently puzzling over as a natural and commonplace part of their education.