Towards Intelligent Databases: or the Database as Historical Archivist

by JAN SUNDIN and IAN WINCHESTER

Databases, as they presently exist, all share more or less the same general character. They are all generated for some specific purpose or class of purposes. And this generation is invariably brought about by laborious manual procedures. The present writers have been involved in the production and use of two historical databases, one in Canada (Winchester) and one in Sweden (Sundin). The Canadian one was generated for the purpose of studying mainly the social structure of a single nineteenth century Canadian city, Hamilton, Ontario. The Swedish one was generated mainly for studying a variety of rural regions in nineteenth century Sweden. The Canadian database used mainly census records and assessment rolls at decade intervals. The Swedish database used mainly parish registers which give a near continuous record of census-like data. Our purpose in this article is to draw upon our experience both to characterize databases of the kind useful to historians as they presently exist as well as to try to characterize what the next generation of databases might be like, namely, intelligent databases.

RESEARCH DATABASES

At the present time, a "database" is a notion associated with computerized data, stored on punch cards, magnetic tapes, discs or other media readable by a computer. One must also distinguish here between a database and computerized data in general. A database, in some minimal sense of that notion, must consist of data organized for a purpose, which can be more or less limited or specified in detail. There should also exist a system producing this organization of data as well as a definite system for exploiting it or using it. Finally, the data and its systems of production and exploitation should be documented in such a way that the entire operation can be understood by someone wishing to reproduce it, use it, or add to it.

Computerized data which does not possess the necessary qualities of purposeful organization, systematic production, systematic means for utilization and adequate documentation should, perhaps, be called a "databank" or a "dataset" or a "datafile," or something else. Such computerized data, and, indeed, non-computerized data, can sometimes with effort be transformed into a database. But it is the systematic embodiment of purpose, production, utilization, and documentation which makes databases the enormously useful things they are for social historians.

There are three general kinds of databases (sometimes, datasets) which are useful for historical studies, particularly of the quantitative social historical kind. The first of these are administrative databases. The second are databases (or more likely, datasets) produced for individual research projects. And the third are databases for general historical research purposes, of the kind which the Demographic Database in Umea represents. We'll say a few words about each of these in turn and then dwell in some detail on the third kind.

Administrative Databases

The traditional source for material for quantitative social historians, historical. sociologists, economic historians and microdemographers has been routinely generated administrative records by government, by business, or by churches. This material has traditionally become available to historical researchers after some legally defined time period, thirty to fifty years in some countries and for some records, one hundred years in others. Almost invariably, the records which have been exploited for historical research have so far been in paper form. However, administrative bodies are rapidly taking advantage of the computer in order to improve their routines or their access to relevant information before planning or taking action.

Computerized data produced and kept by such administrative bodies is potentially as useful to future historians as the routinely generated material of the nineteenth century has proved to be. However, since modern administrations are keeping their data on magnetic tapes (or their equivalent) only, if the present generation of historians and archivists are to have successors, they will have to act to see that such data are preserved.

Administrative data, whether arranged in a database or merely as a dataset, has always posed some problems for the historical researcher even before computerization. The most obvious problem is that administrative data are not usually organized primarily for research, but rather to facilitate some mundane daily task, such as to collect taxes or record local growth of population or determine and list voters. Thus data may be gathered according to no particularly systematic plan and even when the plan of gathering may have been systematic, the execution of it may not have been. Consequently, information of greater interest to the historian may be left out or incompletely gathered. For example, for many years various bureaucracies in countries like Sweden and Canada kept information on country of origin for individuals or for their parents. Recently, however, this has for political reasons been deleted since it has come to be thought to be somehow a blow to one's human rights and no doubt to the historical value of the record. In contrast, for nineteenth century Canada, we know much about how different immigrant groups fared relative to one another and how their children fared. Should country of origin be systematically excluded, such knowledge will be much harder to obtain for future historians.

Not only can there be data left out or incompletely collected in administrative data files, it can be organized in ways which make it very difficult for a historian to use. Although this has always been a difficulty, the solution up to now has been for the historian to sit in an archives somewhere with pencil and paper and record from the administrative records the data which he or she needs. The limitations on the historian's method here is set only by the speed at which he can work and length of his life. One might even argue that for the historian academic freedom, in the form of the freedom to engage in his research, has traditionally been connected with the fact that, in principle, if not always in practice, in a democratic society administrative records and usually all records become generally available after some time limitation. No one is totally excluded from this historical information who is willing to put in the personal effort for the investigation.

The computer changes all this. Like the philosopher and the mathematician, the historian has traditionally needed only pencil, paper and time to get on with his research. The sociologist today, and the historian tomorrow, are very unlikely to find an institution willing to produce sheets of data from its computerized files free of charge. The machine costs money as do the wages of the applications programmer who has to produce the output. And he is likely to have very little time to understand the historian's problem or assist him in getting exactly what he wants. Complicating the matter even further is the fact that codes of privacy or secrecy, especially if two or more data files are involved, are often much more strict for computerized data than for other information. Thus at the moment a sociologist or demographer or contemporary social historian may sometimes find that all these restrictions and limitations make it impossible or unreasonably expensive to get on with his research. Programming time, for example, might be unreasonably expensive, even if one can get access to the data, since the database was not originally organized to meet the researcher's needs. And codes of privacy, both in Sweden and Canada, can be extremely restrictive (except for credit card companies!) when individual data from two separate registers are to be linked together. Indeed, the linking process itself can be very complicated and expensive.

But despite the very real problems which inappropriate organization, high cost of access and difficulty of access pose for the contemporary and future historian wishing to use computerized administrative databases, the most disturbing matter is the easy and systematic erasure of the information. Historians must always work with the loss of records, sometimes by fire, sometimes by age and disintegration and sometimes by administrative stupidity. Sometimes, as with the destruction of the library at Alexandria or Nazi bookburning, there has even been total or systematic destruction of these. But by and large, administrative regimes in the past have wanted to retain the results of their painful handwritten handiwork, as for example the administrative records of churches, particularly the parish registers. But there are lots of other examples as well. Modern computerized administrative regimes, however, have no love for their handiwork (since it is so impersonal and standardized, one supposes). Consequently it has come to be thought convenient to erase totally computerized information once it ceases to fulfill its original purpose. We therefore run the risk of getting administrations "without a history." Similarly, the individuals who are registered by these bodies lose their life histories, at least so far as they could possibly be extracted or assembled by a future historian. Computer erasure, from the vantage point of quantitative social history, is like a black hole into which all information goes and none comes out. The past disappears into the complete computer darkness.

Two examples may be enough to illustrate the impact of this problem. The register of serious criminals in Sweden serves both a controlling purpose for the police as well as providing an opportunity for criminological research. As a research resource, it is virtually the only practical register for longitudinal studies of persons convicted of serious crimes in Sweden today. Both for practical reasons and humanitarian ones, the register is regularly purged of all information relating to a criminal who has died, passed his eightieth birthday, or remained unconvicted during the last ten years. Unbiased study of a particular age group is impossible, especially since there is a relatively high mortality among certain groups of drug abusers. Historians, of course, will have no opportunity whatever to use this register for studies of the past, since eventually all criminals die. Fortunately, the Swedish Bureau of Statistics still keeps magnetic tapes of the content of the register for the most recent years, but there is no guarantee that they will continue to do so voluntarily and with the goodwill of the Data Security authorities.

Another problem is illustrated by the recording of land ownership and use in Sweden today. At the moment, a potentially very useful digitized map of Sweden with coordinates for all estates and buildings in Sweden is near completion. The map is intended to serve administrative needs for land surveying and planning. However, this map is to be accurate only for the actual moment of use. This means that all changes are registered as quickly as possible, and all previous data are purged automatically from the file. The data thus lack any time dimension, and are again useless to historians or historical geographers.

Fortunately, some administrations still keep paper documents which duplicate the kind of information held on magnetic tape or disc. Courts keep records about serious convicts and land surveyors keep archives with documents and maps from the past. The mountains of paper piling up in administrative archives are, however, leading to pressure for new and drastic policies about what should and should not be preserved for the future. In both Canada and Sweden, all new regulations have tended to lessen the requirements for preservation and more and more series of documents are allowed to be thrown away after a certain number of years. As the years go by, the historian may find himself in an increasingly worse situation as he tries to build up the biographies of ordinary individuals from the past in order to do the sort of things which have made quantitative social history such an exciting part of the recent historical landscape.

In case one would like to take comfort in the fact that at least modern data handling methods have not affected the task of working with traditional records, here is a disturbing example. The records created and accumulated by the great Canadian educator, Norman MacKenzie, while he was president of the University of British Columbia were microfilmed and original records destroyed. The microfilming was so badly done that the papers are extremely difficult, if not impossible, to use.

Administrative databases, especially for longitudinal research, can sometimes be a substitute for destroyed documents. They can also potentially help the researcher to get his information much quicker and on a grander scale than he could ever get working manually with paper documents. However, administrative databases must be designed so that historical data is kept intact and made accessible at low cost.

Sometimes this can be achieved, at least partially, without too much effort on the administrator's side. Slight increases in the expense of data storage and retrieval should be worth paying for in order to preserve our social history for our grandchildren.

Contacts and the exchange of views between researchers and at least the most important producers of administrative computerized data could potentially improve the situation. In the case of both Sweden and Canada, our national archives are our natural partner in such contacts, as are our provincial and university archives as well, since they are specially knowledgeable in problems of preservation. In both countries the national archives have already begun the storage of magnetic tapes, indicating that at least the national archives are aware of the problems. Some effort, however, must also be made by the representatives of research bodies and those which fund research in the social sciences and humanities in order that administrative organizations will be aware of the research community's needs, hopes, and fears. Indeed those who feel the need should certainly take the responsibility for initiating the dialogue.

Databases or Datasets Produced for Individual Research Projects

If administrations can be criticized from the vantage point of the historian for not arranging databases conveniently for future studies, the same criticism can be levelled at the historian himself—or at least at his collective self. Most recent social history has featured an individual historian, or perhaps a team of historians, producing a database or a complex of datasets for the purpose of pursuing some single research aim or end. Once that aim has been reached, the database, so painfully built up by manual labour, is usually cast aside, perhaps destroyed.

In some countries there are organizations for the documentation and re-use of databases or datasets produced by and for individual research projects, for example, the Institute for Behavioural Research at York University, Toronto and the Inter-University Consortium at the University of Michigan in the United States. This work is demanding and difficult. The files of data produced by individual research projects are normally made with limited financial resources for a limited purpose. Data are often collected selectively from documents or are the result of interviews designed for the project's specific needs. Often the documentation of the data and the data systems used are unsatisfactory for other users with different intentions from the original ones. As well, data from different sets referring to different groups of individuals or different geographical areas, are often incompatible and cannot be matched in order to increase the number of variables under observation. Not only is there no standardized reporting of data files in the literature, there is also no standardized reporting on how such data files were built up in the first place. This is especially problematical if the files are themselves the result of a systematic process of matching disparate files relating to the same collection of individuals.

Organizations for documentation and re-use can keep a copy of the data themselves, but sometimes the only available version remains with the original producer. In such a case, restrictions on re-use can be very great. The data, while perhaps adequate for the initial research, may be sloppily collected or encoded. Perhaps the original user of the data may wish to restrict the use of the data. Even where the original scholar is willing to share his data with others, his documentation may be incomplete or difficult to understand without special tutoring. This need not be merely bad scholarly practice. Financial limitations put constraints on documentation just as surely as on the collection of data initially or on the research itself. In Sweden, it has been suggested that research foundations should sign a contract with their funded projects to secure a standard high quality documentation for re-use of data produced with their grants. This could be an improvement, especially if the possibility of re-usage is a financial advantage (or scholarly advantage) for other projects applying for money. Such a policy, however, must be used with care. There is no guarantee that such a priority would direct money to the most deserving research projects. And some excellent researchers would shy away from financial support with too many strings attached.

In the case of files, datasets, or databases produced for individual research projects, neither good documentation nor the owner's readiness to allow others to use the data may be enough. Data may have to be organized in a way that fits certain types of output and certain computer configurations better than others. Magnetic tapes may have to be reorganized with attendant high cost, perhaps higher than a re-user may be willing or able to pay. Clearly there is a significant role for organizations concerned about the re-use of research data, who might provide advice to the constructor's of datasets about standard models of data processing, data documentation, and data conversion. Without such advice, potential users and contributers would be rather like a linguist working in a tower of babble. Such organizations could be of very great use for quantitative social history, in spite of the heavy burden of problems they face.

Databases for General Research Purposes

Some of the problems for researchers using administrative data directly or for those re-using data for other projects may be overcome by the construction of databases organized primarily as a service to *all* projects in need of a certain kind of information, particularly where this information is of the routinely generated sort which potentially covers an entire population within a political boundary. Two Swedish examples of such historical databases are the Demographic Database in Umea and the Database for Social History in Stockholm. The Umea database covers a number of parishes and regions in nineteenth century Sweden using the church books as its primary source, while Stockholm database collects data only from the city of Stockholm for the period 1870 to 1910 using the civil nominal registration lists. In Norway, there are two such databases as well, one in Bergen and another in Tromso, each of which registers part of the Norwegian population registers from the nineteenth century.

There are enormous advantages to be had from such large multi-purpose database arrangements. From the vantage point of the individual user, it is as if the administration of a past time had already computerized its data in a way especially convenient for the historian. It is very rare that a single project ever gets the massive funding needed to establish a large, multi-purpose database on the scale we are talking about here, namely, data relating to many hundreds of thousands of individuals, perhaps millions. Yet from the user's vantage point, it is if he had his own project under his own control. While the advantages to the individual researcher are manifest, the advantages to the historical discipline as such are perhaps greater, for many more than a single project can be supported at once. In this way a variety of different questions or aspects of a question can be examined. There are not many examples of this sort of thing having been done successfully as yet on a large scale, although research using the data assembled for the Canadian Social History Project by one of the present authors can illustrate this to some degree. The main user of the data for historical purposes was Michael Katz who was the Project Director from 1968 to 1972. However, a number of doctoral students successfully completed work using the same database on such diverse topics as urban social structure, literacy, historical geography, and school attendance. Similar things can be expected from, for example, the rural history database established by Gerard Bouchard and his colleagues in Chicoutimi. At present the Demographic Database in Umea is working with a wide range of scholars both inside Umea University, and across Sweden, as well as from abroad. We can reasonably hope that it will set the international standard for databases of this type for some time to come.

Large projects of this type must be financed by central funds over a very long time. This kind of commitment can only be made by a determined government or research agency which is convinced that the long-term benefits of such databases outweighs the disadvantages. And the disadvantages are very real. Many years must be spent putting the database in place before there is a research return. In the meantime, the same amount of money could have been used by individual researchers working on smaller projects with nearly immediate return. Any such large database of the general research kind requires special methods and checks at every stage. The transfer of information from original documents must be complete, quite contrary to ordinary historical practice which is selective, because computerized data must stand in for the original documents for as long as the database exists. Not only must the transferral process be complete and foolproof, but the data must be held in the system in such a way that information at the level of individuals can be added freely at any future time from other sources than those chosen as the backbone of the database. The storage and retrieval system must be designed to provide flexible and economic access. The "interface" between user and computer system must be easily managed, and the documentation of the content and methods of input and output must be complete and easily understood.

THE DEMOGRAPHIC DATABASE IN UMEA, SWEDEN

To date the large databases for historical research, which the Umea Demographic Database illustrates, have devised excellent systems for transferral and checking of data. The storage systems and retrieval systems are also highly developed, as are the means for documentation, but there is room for improvement in the interface between the user and the system.

The Umea database began in 1973 as a temporary project financed by the Swedish National Board of Labour in close conjunction with the historical research of Professor Egil Johansson in order to supply work to women in an area of heavy unemployment near the northern border between Finland and Sweden. Using ordinary government funds to finance a very expensive historical project was unique in Sweden at that time. The centre of the project in Haparanda was also an area of high unemployment. Partly as a consequence of this, the number of people employed in the transferring of data from old church records to cards for further processing by machine grew from a handful to twenty-five or so very rapidly. By 1978 progress encouraged the authorities to give the project a permanent status within the University of Umea.

Swedish and Finnish population registers from the nineteenth century are a unique source of information. Nothing like them is to be found anywhere else in the world. They cover the whole century with the registration of births, marriages and deaths and also migrations. These records, along with the literacy and religious knowledge examination registers, give a continuous record of each resident of Sweden during his or her entire life. Indeed, a similar registration system was used by Swedish-American congregations in North America during the second half of the nineteenth century, making it theoretically possible to document Swedish emigration. Entire populations can be observed or a single individual can be followed continuously throughout his entire life, in striking contrast with the periodic appearances of a name in parish registers, tax rolls or the census in other countries.

The special quality of these registers were an important argument for initiating the project in Sweden. The possibilities which these registers appeared to offer were especially interesting once Louis Henry in France and the Cambridge Group in England had demonstrated the value of so-called family reconstitution for the study of historial demography. In the Swedish case, a great deal of the family reconstitution was already done by the manner in which the registers were set up. The reconstruction of the life-histories of individuals and families proved to be valuable for the history of social and geographical mobility, the social history of children, women, the aged and special occupational groups. Such individual and family life-histories could also be used to study land use, trends in social behaviour and the history of medicine, including the study of hereditary diseases. In addition, population registers can be used to link individuals represented in taxation poll registers, prison records, lists of workers in factories and mills, inventories of deceased persons and school registers. Because of their comprehensiveness and completeness parish registers provide the basic structure to which data from other registers and records can be added. For a number of Swedish parishes and regions, once the population has been reconstructed in a database from the parish registers, information can be added as it becomes available from other sources. The fact that the original sources become more fragile each year due both to age and intensive use, led the Swedish National Archives to support the project strongly. Indeed, the National Archives acted as the administrative head of the Demographic Database until it became independent and permanent in 1978. From the vantage point of the National Archives, a database would diminish the use of the original sources and increase the probability that they would be preserved in the long run.

According to the government regulations establishing the Database, it should "register and prepare mainly demographic and social historical data for research, education and archives and make data available for research." It should also "work for research cooperation within the fields where data finds its use and, when appropriate, take part in methodological development." Five principles were initially agreed upon by the steering committee for the Database, namely:

1. The data should be faithful to its sources (i.e., one should be able to compare in detail the registered data with the source data).
2. The dataset should be complete (e.g., a complete rather than a partial set of parish registers should be registered).
3. The database should be open in its organizational plan (i.e., it should be possible to add to the database easily and in a variety of ways, for example, over an extended time period, within greater geographical areas and with different sources).
4. The data should be prepared in a way that makes both genealogical and cross-sectional demographic research possible. It should be possible to analyze the total population or a defined sub-group for both a single year or for a particular time period.
5. Data processing and access should be inexpensive.

These principles may appear innocent and obvious. However, like any set of principles, they may be very difficult to adhere to in practice—especially if they should turn out to be mutually incompatible. There is also a principle, only slowly dawning on us, that we would like to include along with the previous five at some future data, namely:

6. The database, as an access system, should actively aid the researcher in his quest for historical, demographic, or genealogical understanding.

This last principle, which will be the basis for our discussion concerning the possibility of intelligent databases in the final section, arises out of our experience with users of the database. However, before we consider it, we shall turn to a discussion of the first five principles.

The question of *fidelity to the sources* is apparently quite simple. Whatever is transcribed from the original sources into the database record system should be the same as what was on those sources. Nothing should be changed. No constructions should occur on the route from the original source via the database to the researcher. Given, however, that (at least in the case of the Umea database) the process of transcription involves many human steps by dozens of people, unintended errors creep in. Thus, fidelity to the sources cannot be an absolute demand. Rather, it must be a statistical one. In the case of the Umea base, the procedure to assure fidelity is to sample the records in a particular batch, comparing each record in the sample against the original sources. The batch will be accepted only if the number of errors on each variable is below a predetermined value. If not, all records are registered again and a new sample is taken for a new check.

For a researcher, however, statistical fidelity may not be satisfactory. Because errors are always present to some degree, a researcher may discover some of these when looking over his data file. Sometimes, especially when multiple records relating to the same individual are used, errors can be detected and corrected by processes of logical reasoning. Indeed, for research purposes, the "errors" detected may be both due to the transcription process and the original recording process: the parish priest sometimes made errors too.

An example of this would be as follows: we wish to produce basic statistics about a population's age structure. What do we do if for some individuals we lack information concerning their day and month of birth? Should we omit such individuals and so, perhaps, bias our data? Or should we use other information, say their date of baptism, and so construct a date for them? Our practice shows that for nearly every individual registered there will be at least one item of information for which the data are either ambiguous or lacking in at least one source file, but in which logic indicates the correction information.

One could, of course, choose to accept the sources as they are, and leave the decision for the handling of such ambiguous or missing but correctible data to the researcher. When two records tell two different stories, a researcher might be left himself to systematically choose in all such cases in his research file. However, this can be a very time-consuming matter for the researcher and for programmers. Many users would willingly accept some construction in a standardized manner as long as it is documented. The Demographic Database has chosen this latter strategy, but keeps at the same time so-called "source files" where the original source data are as unchanged as possible. Such a strategy requires a great deal of effort from the database staff in defining and applying the principles of construction in order to produce "most probable life-histories" for each individual. But it saves the same job being done over and over again by each individual researcher in a variety of ways and with different standards.

Because this strategy has been in fact chosen, the researcher can compare the "standardized" life-history with the testimony of the original sources. In theory, one could offer the researcher a note concerning each standardized choice whenever there has been a systematic construction as well as a pointer to each volume and page in the sources from which the life-history has been constructed. This would increase the cost of file storage and production considerably and would increase the volume of the master file considerably. The Demographic Database has, in fact, chosen to document the principles and priorities of construction. As well, the user can get a list of the places where his individuals are recorded in the original sources. He can, for example, compare a sample of the proposed life-histories with the original sources in order to make up his own mind and modify the standardized judgments should he so wish. He can also ask for a file according to other principles provided he can pay for it.

The principle of *completeness of data registration* is not difficult to understand as an aim. Nor is it methodologically difficult. Every notation in the available sources ought to be registered and available for users. However, for such abundant sources as the Swedish Church Registers, this requirement is both very expensive and time-consuming. Even though about thirty people are now involved in the input process at the "production unit," it takes about a year to add another two to three parishes to the database. Since there are about twenty-five hundred parishes to be registered, one thousand years would be required for a complete registration. There are only two ways of speeding up the process. One would be to neglect part of the available information. The other would be to add personnel to the production unit. The first strategy has been rejected because of the impossibility of foreseeing future research interests. The second is a function of a nation state's ability to pay and its commitment to historical research. Clearly for the remainder of this century the Demographic Database will have to select the parishes which it registers completely according to a scheme which is maximally useful to historical research, since it cannot register them all. For this purpose a research steering committee functions in close conjunction with the Database, selecting priority parishes or regions for research. However, like the neglect of data, the selection of parishes *a priori* suffers from the inability to predict future research trends and needs.

Satisfying the principle of ease of *access to the data* means, with present technology, expensive storage on permanent computer discs. But all the data from even a single parish takes several megabytes to store. Perhaps new generations of computers will have the capability of storing almost unlimited volumes of data for immediate access. But for the moment it is not good economy to keep all the information on permanent files. The Database presently keeps only some of the life-history information for all individuals registered so far on permanent disc files. The remainder of the information must be taken from the so-called source files or from other magnetic tape files. It involves extra cost to transfer data from one medium of storage to another. But access, though slower than permanent disc storage, is always possible. Clearly the present system is a compromise between quick access and economy of storage. A similar compromise has to be made in any institution which wishes to create a large database but which has limited computer space.

For research purposes, perhaps the most important reason for maintaining such an expensive database as the Umea one is its being *open to additions* of other information pertaining to the individuals in its master file and thereby available as a participant in an unlimited variety of research projects. Practically, this means that there must be easy access to the identifying items related to each individual in the master file. Without this condition being met it would be impossible to link new records at the individual level to old ones already in the file. Such linkages can be accomplished by hand. However, since the files are so large this is impractical. At the present time the Demographic Database is planning a general linkage system which would enable the addition of new files at will. In order to accomplish this, the characteristics of the identifying items in the master files must be studied systematically and various standardizing strategies devised. One example of this is the standardization of the spelling of surnames so that candidates for linkage are sorted initially into a most likely group or "pocket" of individuals for further detailed comparison of such linkage variables as place and date of birth, age, identity of parents and the like.

With special programming, depending upon the nature of the research topic, it is possible as the Database is presently organized to have *both geneaological data or demographic data.* However, were it to be economically sound at some future date, one would ideally like to have the master file contain all the appropriate information organized in an appropriate list structure so that building up family groupings, individual groupings, or inter-generational groupings would be a simple matter of standard programming via some interactive mode or other.

The previous discussion points to the necessity of a variety of tradeoffs in attempting to fulfill our fifth principle, namely, *cheap data processing and access.* Thus, fidelity of the sources has required us to compromise and to produce constructed life-histories. Completeness of the dataset has meant that we must select only certain parishes for immediate registration. Openness of the files to future additions, while crucially important for future research, has had to be treated by us as a matter for future efforts in terms of name standardization analysis and general record linkage programs, since it is expensive. And instead of an ideal list organized into a master file available on disc we have a compromise arrangement with only some of our data available for immediate access. In many respects, then, the Demographic Database in Umea, while a wonderful research tool of unparallelled magnitude, is far from an ideal database.

Comparing The Swedish Demographic Database With An "Ideal" Database For European Or North American Data

The nineteenth century in both Europe and America saw a series of parallel attempts by governments to keep much better track of their populations. Educational and social reforms were the primary motivation in nearly all cases. In this regard Sweden was one of the leaders. So was Canada. It is therefore instructive to compare the actual Demographic Database in Umea with what one could envision for Europe and North America as being an ideal database for the nineteenth century. Such an ideal database would still lack the qualities of our suggested sixth principle of organization for a database, namely, that it *actively aid the researcher* in his pursuit of historical understanding. But it would be the basis on which such an intelligent database could be constructed.

An ideal database, if cost were no object, would look much like the Demographic Database without compromises on the first four principles and with the addition of other available information. Although a historical individual needs not be a human individual, historical individuals such as "the people of a town during the century," "the workers in the iron mines in 1850-1860," "the families of the parish with one or more children" and the like are all composed of human individuals selected in certain ways. Consequently, the backbone of any nineteenth century "ideal" demographic database must be some basic data source which registers the entire population of a country over a particular time period. In this regard, the data source for the Umea database is as near to the ideal as possible for the nineteenth century. In most other countries parish registers were neither comprehensive nor continuous. Migration tended to make their value questionable. Also census taking, even when it became universal in the mid-century, was always a less than complete enterprise. In essence, the census is a snapshot of a particular country at a particular time rather than a motion picture of that country over the entire century. The Swedish Church Registers, in effect, give us the motion picture. Canadian censuses, by contrast, give us snapshots beginning with 1840 for each decade, and the initial attempts are by no means high quality records of their times.

Besides providing a backbone register of every person during the time period in a particular country, an ideal database would also possess files of all the other data available in archives (or other sources) which pertain to the individuals in the backbone. It would, that is, possess a complete biographical characterization for every individual in its backbone file. The nineteenth century, both in Europe and

America, produced enormous numbers of documents pertaining to ordinary individuals. Thus, there is taxation data, data relating to property possession and registration, data relating to schooling and school attendance, data relating to the legal order, including criminal data. The total amount of information which could, at least in principle, be related to the backbone file is this very great. This is perhaps the aspect of the Demographic Database which is at present farthest from the ideal. Although the design of the Database certainly permits the addition of new individual data, such addition would be a matter of special programming each time it were to be added. In an ideal database, there would be an easily updated system of organization for the master file such that any new individual data would be located in its logical place without the necessity of the disruption of the entire master file arrangement. But this in turn means that the master file and the backup system software would recognize the importance of record linkage as a standard part of the database management and upkeep. Perhaps the computerized files of the Mormon Church in Utah are nearer the ideal in this regard than any other at the moment.

Since a database is not just a permanent record of social historical documents but a research tool, an ideal database would maintain a record of all the sub-files which were generated using it as source, as well as a record of the research results to date. Much of the research results could function as qualities to be attributed to higher order historical individuals. This suggests that an ideal database would also be organized such that not only could data relating to individual persons be added, but also information relating to virtually any conceivable level of historical individuality which the files permit. Since this would be unlimited, in practice some initial structure for such historical individuals would have to be specified in advance (for example, definitions of "family," "household," "houseful" would have to be produced). Here, the closest we are likely to come to the ideal would be in devising a very flexible file organization system initially such that virtually any grouping of individuals would be possible to produce with a minimum of subsequent special programming.

One would like, for purposes of any demographic research project, to be able to compare any arbitrarily chosen local region with both a country as a whole or with any other local region. In the case of the Demographic Database the local regions are dictated in advance by means of the steering committee which chooses the order in which parishes are to be registered and the regions from which they are to be chosen. Since it would take one thousand years at the present registration rate to re-do the nineteenth century's registration efforts for the Church Registers alone, there is no comparison with the data of the country as a whole available either, except for a few categories assembled in the nineteenth century as well.

This leads us to a feature which any database should have and which an ideal one would have in full measure, namely, complete aggregate statistics for the main variables relating to the backbone files. Easily available aggregated statistics concerning all administrative units exist for most western countries in the last century. In Sweden this kind of information is relatively abundant. For example, after 1749 the Swedish clergy were required to return statistical tables concerning births, marriages and deaths each year to a central agency. For every third year between 1750 and 1800 and every fifth year between 1800 and 1860 and every decade thereafter, there are census-like tables concerning the demographic composition of the population. These statistics have been aggregated at the deanery and county levels. At the county level the statistics are usually published, but for lower levels they are available only in the original form kept in archives. The Demographic Database has registered only the tables for 1806-07 and 1855-56 so far as an aid to our researchers. But future plans are for much more of these to be routinely included in the data available to the researcher. Ideally a database would have all such aggregate data registered as well as the various nineteenth century census statistics, so common in the latter half of that century, not only concerning families, occupations, school attendance, literacy and the like, but concerning agriculture and manufacturing as well.

An ideal database would be able to pursue longitudinal studies as far as they lead. This means being able to trace individuals or families, for example, throughout a complete series of physical movements from place to place. With a regional database, as the Umea one will remain for a long time, such tracing, while in principle possible, is in practice limited to the population which remained within the region during the time period of interest.

If the research at hand cannot be satisfied with the study of a population within a given region, then the database as it is presently organized cannot do the whole job. Special work by researchers tracing families and individuals through the church books by hand would be necessary. Perhaps, for those who would basically be satisfied with the study of the people originating from within a specific region, a few sample model cohorts could be traced. Or one might simply study the character of migrants compared to non-migrants. Though in this case we would still have a selection of Swedes who lived part of their lives in a particular geographical level.

Even if the Demographic Database did pursue certain cohorts of migrants throughout the country during the century, it would be unlikely to satisfy a researcher's needs for the pursuit of particular migrants with particular characteristics. Nor would it be cheap, since the sources which would permit this are scattered in a variety of archives. Thus for the following of internal migrations, the Demographic Database is far from ideal as well. However, within fairly large and well-defined regions, it is possible, probably for the first time anywhere, to follow the total migration patterns of enormous numbers of people who happened to live their entire lives within the region. It may not be ideal. But it is better than anything that has been available hitherto.

One of the difficulties at the present time with all historical databases which are built up like the Demographic Database around a routinely generated backbone of data is the relating of that data to material which would be more naturally stored on paper documents or, perhaps, on microfiche. Here it is hard to specify in advance what the ideal would be. We imagine that microfiche and its future analogues constitute an area which those interested in building databases should watch carefully to see that such data storage is compatible with database storage schemes. Perhaps the sort of thing we look for here as an ideal is a scheme of cross-referencing between databases and microfiche archives and traditional archives such that an integrated system of data and information available for research is produced which would be independent of the medium of storage. For this to come about, and for the specification of an ideal here, a close relationship and active cooperation between databases, traditional archives, and researchers will be necessary.

Perhaps the greatest single gap between any actual database, including the Umea one, and an ideal is the lack of international comparison both in terms of organizational standards as well as in terms of basic statistics, comparability of data and research results. In this sense, an ideal database would not have a national basis. Rather, it would be international from the start. It is beyond even the largest and most well funded database to follow international developments and to engage in cooperation between database projects and researchers in several countries. This has been felt by the Demographic Database in Umea and it was felt by the Canadian Social History Project even when it was engaging in collaboration on the Five Cities Project in the early 1970s.

The problem of comparability and exchange of data is both a technical and a conceptual problem. In its technical aspects it could probably be handled by international cooperation between organizations for the documentation and re-use of research data as well as by major databases. However, the conceptual problems, including problems of validity, depend upon sub-groups of researchers meeting in international settings to discuss their specific problems in their specific research specialties. We do not have enough international conferences specifically aimed at such matters.

In summary, then, the notion of an "ideal database," while an attractive one prior to trying to conceive of what it might be like, does not seem to be such a helpful notion when confronted with the varieties of historical need, the extent of costs, or the organizational difficulties. There is no single database strategy for all possible historical problems, even if one restricts consideration to questions of the maintenance of data on the individual level. In different countries the most appropriate databases for historical research are likely to differ depending upon the sources available. Different sources dictate different research strategies. And different research strategies lead to different database design. There is, to our knowledge at least, no general database strategy which will satisfy all research needs. Thus, for example, while the Umea database is well fitted for studies of geographical regions, it is not suited for studies involving national samples or national averages. Of course, if it were possible to have a few thousand people working on the transferral of information we could record all the data available and so generate national data.

In practical terms, then, any actual database is likely to fall short of a conceivable "ideal" because of the necessity of engaging in compromise solutions to the support of a variety of principles of organization, each worthy on its own, but mutually incompatible either for conceptual or cost reasons.

However, although actual databases may not be ideal databases, they can be improved, we believe, in one very significant way, namely, in their relationship to the users. It is to this topic we wish finally to turn.

Towards An Artificially Intelligent Database

Databases, as they presently exist, all share more or less the same general characteristics. They are all generated for some specific purpose or class of purposes by laborious manual procedures. In the normal case, their organization is such that the data which they contain can be had only by some simple, direct means of access or else by elaborate and time-consuming (somtimes even heroic) programming measures. The data-bases which serve for airlines, for example, are available at a variety of terminals at airports and at other ticketing centres. However, the questions which can be asked of the database are strictly limited. The system is designed for maximum efficiency in answering a restricted number of questions. Should an airline executive wish information which the files potentially contain but which cannot be had from the precise question and answer sequence designed for the ticketing terminals, he must let a programmer who understands what he wants go away for a while to devise a way of extracting the information. Depending on the nature of the information and the organization of the files, this can take from hours to months.

The situation is not much different at present even in research databases. In the standard case, the files are organized according to some well-documented and strict plan. Some of these files, or perhaps a portion of a master file, or perhaps a "show" file, is available directly through a remote terminal on which a limited variety of data display can be arranged. Perhaps even a little data analysis of a limited kind is available through a remote terminal too. However, all of this is for familiarizing the researcher with the resources of the database and permitting him some sense of what he might do with the material in the database. When the researcher has an idea of what the database contains and of what he needs to help him handle his historical problems, he must then plan some means of extracting the necessary data from the database file. Sometimes he can do this by means of such longstanding analysis packages as Datatext of SPSS or a host of others. Sometimes, however, the analysis he wishes requires very elaborate special programming and, perhaps, statistical analysis. Here there are usually three main programming steps. First, a special file of data has to be created for the researcher. Then special tabulations from that data have to be created. And finally, if there is to be a special statistical analysis, the tabulations must be processed in some appropriate statistical way.

In all of this, the researcher is active and the database passive. All of the intelligence is supplied, at least in the standard case, by the researcher. Except for certain medical-diagnostic examples, there has been virtually no application of the techniques of artificial intelligence to the database.

What would an intelligent database look like? Roughly, the intelligent database would be a combination of historical archivist, systems analyst and programmer, co-historian and friend.

To give some sense of what we mean, here is an example of a conversation between a historian and an intelligent database, perhaps in Umea, some time in the near future.

Historian: Hello, Database. I am the Swedish historian Borje Salming, from the University of Maple Leaf, in Toronto, Canada.

Database: Hello, Professor Salming. I am the Umea Demographic Database. How may I help you in your research?

Historian: Well, I am especially interested in Swedish cities in the nineteenth century.

Database: Ahhh! Then you may be interested in four cities we have registered to date. We have complete church register data and other back-up data beginning in 1807 with continuous updating until 1910. The cities are Haparanda, Sundsvall, Uppsala and Gothenburg.

Historian: Yes, indeed I am. I'm also interested in migrations from and to cities in the nineteenth century.

Database: I've got all the dates of in- and out-migration for the four cities, to all other parishes in Sweden as well as personal and family data on each migrant for the entire century. Would you like to see some national, regional and city summary statistics on migration? I have also sex and age breakdowns. But family size would take me some time.

Historian: Yes, please.

This is the sort of conversation which any historian has had at some time or other with a helpful archivist or with a co-researcher. The archivist or co-researcher, however, even though very helpful can at best aid the researcher in orientation. A database capable of such a conversation and in possession of the data in an accessible form could be an active aid to the scholar.

There are two questions which need to be considered with respect to the possibility of an intelligent database. The first is whether we have the programming and data processing capability today to devise such a database. And the second is what the logical components of such a system would be. As regards the first question, there is every reason to believe that the programming techniques exist, since the conversational programs for medical diagnosis are now highly developed. There is still some question about the direct access memory capability of machines available now and in the future. However, recent technical advances suggest that these can be very many times faster of access and larger in memory in the next generation of computers. Assuming, therefore, that in terms of programming, machine memory, and machine speed there are no problems in principle, what would the database systems analyst have to consider to be the basic components of such a system? In order to determine this, the systems analyst would have to know what a researcher would require from an intelligent database. We would suggest the following:

1. An intelligent database would be able to converse with a researcher in a natural language concerning the material in the database and concerning all of the plausible types of analyses in which it might be expected to be involved.
2. An intelligent database would have to take hints and suggestions from a researcher and offer plausible interpretations of these in terms of data lists, tabulations or analyses.
3. An intelligent database would be capable of re-programming itself to perform new analytical tasks as required by the researcher.
4. An intelligent database would offer interpretations of the data it has already processed and engage in discussion with the researcher about the interpretation of the data.
5. An intelligent database would actively aid in the co-authorship of papers.

One is tempted to suggest that it ought also to be able to brew coffee and order tickets for the ballet as well as babysit the kids. However, in our vision of an artificially intelligent database, what we have characterized as the five main activities are activities parallel to what a database like the Umea one actually performs at the moment with an enormous expenditure of human time and personal involvement. In this sense, then, the Umea Database (conceived as a system which includes data files and tapes, a means for the transferral of data in a systematic way, a system for the processing of the transferred data so that files are continually updated and maintained, and a variety of people involved in administrative, transferral, programming and systems analytic tasks as well as statistical ones) is an intelligent database. What we want in an artificially Intelligent Database is an entirely computerized, interactive database which accomplishes directly in a reasonably short space of time in interaction with the researcher what this vast apparatus of people and things presently does over a longer time.

Whatever system plan a systems analyst would devise for setting up an intelligent database as we imagine it, there would have to be, it seems, the following components: first, a natural language conversational system; second, a data-retrieval and analysis system related to that conversational system; third, an internal programming system tied to the conversational system; and fourth, crucially, a data-organization scheme such that indefinitely varied access to the database's data is possible. The natural language conversational system would have at least four sub-components: (a) a conversational mode related to acquainting a researcher with the resources of the database; (b) a conversational mode related to the discussion of possible data analyses and the production of these; (c) a conversational mode related to the re-programming system; and (d) a conversational mode related to data-interpretation and the writing up of the research. Artificial intelligence has explored to a considerable extent the problems connected with the first of these sub-components. It would be relatively straightforward to devise a good conversational program to acquaint a researcher with the resources of the database. Devising a conversational program relating to the discussion of possible data analyses and the production of these would not be very much more difficult than the first. In both cases the major problem is the anticipation of needs and requests and the capacity to detect and respond to key words in appropriate and meaningful ways.

The sub-component which would be a conversational mode programming system would be much more tricky. The problem here is that of devising a general purpose data-analysis programming system which can respond to typewritten or oral commands in a precise way, even when the commands may themselves be vague. This is typically the situation of the applications programmer in conversation with a researcher who now has some notion of the files in a database and a more or less clear idea of what analyses he would like.

Finally the "discussion of results" sub-component would be very difficult indeed to devise. The problem here is devising something which, like a researcher with years of experience, can make some sense in a coherent way of what his data-tabulations and analyses really mean. This would be tantamount to devising not merely an artificially intelligent program that has a fixed repertoire of responses, but one which in some sense *understands* what it is doing as well.

It may be that even in this latter case, a wide repertoire of responses will be all that we can ever devise to stand proxy for what understanding is for us. Certainly one component of our understanding is our being able, in a given set of circumstances, to take appropriate action. But taking action appropriate to the circumstances is not the same thing as taking action circumscribed by the circumstances. It is this difference that a truly intelligent database would somehow have to embody. The difference is a well-known one in the theory of language. Noam Chomsky pointed out some time ago that the creative aspect in language use as such is just this capacity to produce new sentences appropriate to a given set of circumstances while not being in any sense bound by those circumstances. While this seems to be true, if true, it points to a mystery which amounts to a stumbling block for the would-be deviser of an artificially intelligent database.

All of the above pieces of complicated planning and programming would be related to the conversational mode of the intelligent database in its relations to introducing the database, retrieving and analyzing the date, devising new and appopriate programs to produce new data analyses or kinds of these, and finally interpreting such data. However, outside of the conversational mode which acts as a control and key by means of which the nuts and bolts of the programming is brought into play, one would need the nuts and bolts as well. Here, while some of the programming might have to involve new and interesting principles, there doesn't seem to be any *a priori* stumbling blocks such as the problems of understanding seems to pose.

The data retrieval and analysis system could be any one of a number of well-known variants. Obviously there are certain standard kinds of questions which any social historical data retrieval and analysis system must be able to handle: n-way cross-tabulations, the making of special lists, the selection of particular individuals or particular groups of individuals, the production of factor analyses, regressions, chi-squares, and so on. Present day analysis packages are primarily for data code in some standard form compatible with the analysis package in question. What would be challenging for the designer of an artificially intelligent database would be the devising of a data retrieval and analysis package which would be flexible enough to both generate any desired sub-file for "eye-ball" inspection, any desired tabulation, and at the same time possess the capability of turning the sub-files into convenient logical arrangements for the variety of standard analyses of use to the researcher.

The internal programming system would have to have the capacity to respond to general instructions of the sort which a historian might give to an applications programmer with a considerable grasp of the needs of the historian. It would presumably have a complex of pre-programmed sub-units which could be joined together in a variety of sequences and inter-relations much the way in which a vocabulary can be joined in a natural language by means of a grammar. We realize that here we are rather like Francis Bacon characterizing the future of science, since we cannot see more than very dimly through a dark and fogged glass. Perhaps, though, a bit of dialogue might illustrate what we are suggesting here.

Historian: I can see that the population of the four cities was very different in the year 1900 from what one would have expected in 1850. Could you project what, had the growth rate remained constant from the 1850s, the populations would have been in 1900.

Database: Certainly. Just give me a moment to adjust my programming. I'll assume a liner extrapolation and compensate for the growth and loss of population by catastrophic death and migration. Yes, here's what the tables would look like. . .

Historian: Just as I thought. Swedish cities would have looked like German cities if there hadn't been such a staggering net migration loss during that period to North America.

This sketches, but certainly doesn't define, what we mean by an internal programming system which would respond to conversational suggestions and requests in the manner of a human applications programmer.

The final bit of programming necessary for our intelligent database is that connected with the requirement for a data-organization scheme such that an indefinitely varied access to the database's data would be possible. In practice this means that various orderings of the data which are necessary in building up the master file (or files) in the first place does not restrict the possibility of re-orderings too severely. For nineteenth century data, for example, of the sort to be found in the Umea Database, the natural orderings are in terms of individuals, family groupings and household groupings, as well as village groupings, towns, counties and cities. In this sense natural means "reflecting the manner in which the data were found organized in the original sources."

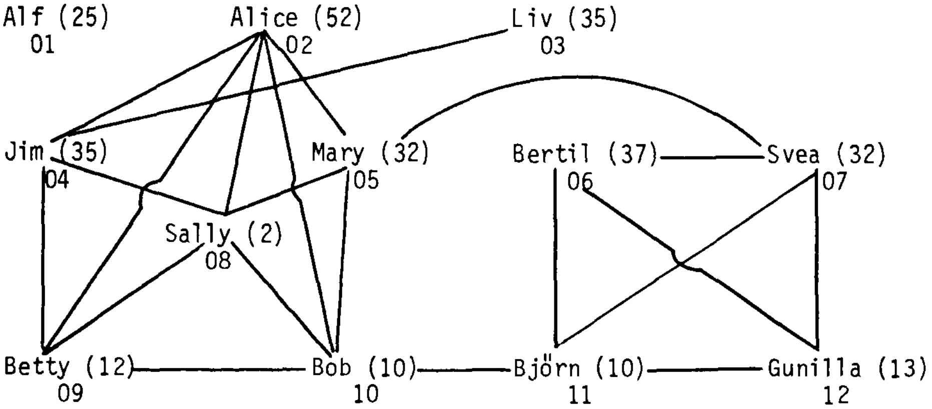
The problem for devising the data-organization scheme is that of, on the one hand, leaving the original orderings logically intact, while enabling new data to be added at will without disturbing the initial ordering minimally. On the other hand, the organization scheme must permit the easy logical re-ordering of the data so that other than the "natural" groupings can be structured at will.

One might, for example, conceive of a data-structure such that any individuals who have "something in common" are logically linked to one another by a cross-referencing scheme. Perhaps we can illustrate this by imagining two families, the Smiths and the Svensons, who live in the same village, say, Stockholm, Ontario. The Smiths, Mary and Jim, have three children, Betty, Bob and Sally. In their house lives a grandmother (Mary's mother) and a border, Alf.

In the Svenson household live the Svensons, Svea and Bertil, and their children, Gunilla and Björn. They have a border, Liv Ullman. Let us also imagine that other than their family or household relationships, all we know of these people is their age and sex. We can list the two families as follows:

The Smith Family		**The Svenson Family**	
Jim Smith (father)	35	Bertil Svenson	37
Mary Smith (née Brown, mother)	32	Svea Svenson (née Borg, mother)	32
Betty Smith (daughter)	12	Gunilla Svenson (daughter)	13
Bob Smith (son)	10	Björn Svenson (son)	10
Sally Smith (daughter)	2	Liv Ullman	35
Alice Brown (Mary's mother)	52		
Alf Applebee (border)	25		

If our ordering principle were to establish a logical link between all cases of individuals have "something in common," then we have at least the following possibilities for our two families: "belonging to the same household," "belonging to the same family," "being a father," "being a mother," "being a son," "being a daughter," "being a grandmother," "being a border," "being male," "being female," "having the same age." Thus we could diagram their various relationships by means of a graphical structure as follows:



We draw in the various relationships (in this case assumed to be symmetrical) by means of a different colour coded for a different relationship, or a different dotted line. Such structures can be conveniently represented in data processing terms by means of graphical codes representing codes for individuals and codes for relationships which both specify the kind of relationship and the address of the next individual satisfying it.

Let us consider only the relationships "same family," "same age," which we have coded as blue and red respectively. If we start with Alf and ask for all those who belong to the same family, the list stops with him. But if we start with Alice, then we must follow through all the blues directly connected to her, namely, Jim, Mary, Sally, Betty and Bob. If we ask for all those with the same age, starting with Jim, we get Jim and Liv. And if we begin with Mary we get Mary and Svea.

In this way we could devise a structure which would permit very flexible tracing of individuals and all their characteristics through a file. Such a file structure, in a more complicated form, would be necessary for the flexible questioning and answering we envisage for an artificially intelligent database.

Conclusion

We can summarize the discussions in the following pages very briefly. There are three basic kinds of databases presently extant which are of potential use to quantitative historians and their ilk: administrative databases, databases specializing in the re-use of data already use for some special historical purpose, and general purpose databases. The Umea Demographic Database is a good, perhaps unique, example of the third kind. Although this database is very useful as it stands, it has a variety of intrinsic limitations, making it less than an ideal database. However, when one envisages the various characteristics of an ideal database there seems little chance, short of infinite funds and infinitely fast computers with unlimited storage, of our ever possessing such a thing.

However, for a real database, such as the Umea one, we can imagine a series of steps towards a database which would actively aid the historian in his researchers. We have attempted to characterize in a general way how one might bring to bear the various techniques of artificial intelligence into the consideration of databases, by specifying what the general characteristics of such an intelligent database might be.