History and the Data Archives

by Hans Jørgen Marker[1]
Danish Data Archives

The Usage of Computers in History in Europe

The Mainframe Age

In Europe the usage of computers in historical research has risen sharply in the past few years. This clearly has to do with the increasing availability of computing power in the form of micro computers. Although a hard core of computer buffs in the field of history completed large projects back in the 60's and 70's it was both in numbers and methodology a small minority in the historical sciences.

The very first project applying computers in history was the "Index Thomesticus" by Padre Busa, dealing with the writings of Thomas Aquinas[2] This project was in the field of indexing source material. Shortly afterwards another major trend emerged, the coding of the contents of sources. Sources were standardized in a way that facilitated computerized analysis especially by means of statistical software or rather crude business type database software.

Among the major database projects of the mainframe computer age are the large Swedish demographical databases of Umeå/Haparanda and Stockholm. These projects are too large to be considered typical, but are on the other hand to some extent only larger scale applications of principles which governed most of the historical computing of the time.

The projects of the earlier days usually relied on one or several systems programmers being employed on the project. The software developed for the projects was dedicated to the project in question and though a number of roughly similar database management systems were developed they were almost never used by other projects.

In many cases the developing project was most willing to share its software with others, but the performers of new projects felt more inclined to develop their own systems rather than to adopt something existing.

Among the exceptions from this rule is the CLIO–system developed for the Sperry UNIVAC at the Max Planck Institut für Geschichte in Göttingen. This system has been used with success in connection with 20 to 40

[1]Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Washington, D.C., May 26–29, 1988

[2]Manfred Thaller: Data Bases versus Critical Editions, in Data Base Oriented Source Editions, Papers from two Sessions at the 23rd International Congress on Medieval Studies, Kalamazoo, Michigan, 5 - 8 May, 1988, p.2.

research projects.[3]

The PC Age

With the proliferation of PC's since the beginning of the 1980's, the general pattern of the application of computers among historians has changed fundamentally. The audience for historical computing has increased dramatically, not only has computing power become more abundant, friendly and cheap, it also seems that many historians feel more comfortable with having the computer on their desk than in the building next door.

Naturally, the first thing which aroused the enthusiasm of the historians, when they got their hands on PC's, were the possibilities in the field of wordprocessing. This enthusiasm has not yet died out. Historians and other people of the arts departments are zealously discussing the advantages of one package over the other. Even at scientific conferences, you may still come across papers praising a particular wordprocessor as the ultimate research tool. Another wing of this discussion is the debate on how to apply the text formatting packages of the mainframe computer days in the micro computer scene, and on the methodological advantages of text formatting as opposed to mere wordprocessing.

Later on the use of commercial database management systems has caught up. The most widespread use of these systems is the making of computerized files to replace the cardbox files which many historians have kept from time immemorial.

Also some historians have become programmers as historical computing is now being done on an individual basis more than within the framework of large-scale projects. The historians' programming effort follows several different paths, such as the presentation of historical data for educational purposes or the making of tools for solving specific research questions. A major area of historical program development is record linkage, a question which has always attracted a lot of attention, also in the days of mainframe projects.

The very large diversity between the uses of computers in history was seen at the two conferences of the Association for History and Computing 1986 and 1987. The proceedings of the 1987 conference has been published[4] and the proceedings of the 1987 conference will be available soon.

Historical Informatics

A specialized theory for the application of computerized methods in history is beginning to emerge[5]. This particular branch of science may appropriately be called historical informatics.

The general concept of this trend in science is that there are questions in history which are specific to historical science and which demand special software solutions, which are not likely to be provided by commercial software developers or by any other nonhistorians for that matter.

[3]Manfred Thaller: KLEIO, Ein fachspezifisches Datenbanksystem für die Historischen Wissenschaften, Göttingen 1987. KLEIO is the PC version of CLIO.

[4]Peter Denley and Deian Hopkin (eds.): History and Computing, Manchester 1987.

[5]A justification of this specialized theory can be found in Manfred Thaller: Why do we need a Theory of Historical Computing. In "Proceedings from the second annual Conference of the Association for History and Computing," edited by Charles Harvey, Manchester 1988.

The questions that a historian would like to put to a given data collection are usually not of the kind that is supported by searches for a specific word in a given field or for some numerical value in another field.

Many terms and objects are time dependent and dependent on context. By making time a part of the context they could be referred to with the general term of "context sensitivity". An obvious example of time dependency is the exchange rates of the different coins of the Kipper Wipper time[6]. In Denmark the number of 'skilling's to a 'daler' increased from 64 to 96 over a period of 23 years. At the same time, the monetary units of that time in Denmark are an example of dependency of context as such, as the term a 'daler' in an official document of, say, 1611 would mean a 'rigsdaler' of 74 'skilling's, while in a private estate ledger or letter it would mean a 'sletdaler' of 64 'skilling's. Another obvious example of time dependency would be the boundaries of a specific territory, say Prussia.[7]. A third example of context sensitivity is the term a hundred, which in north western Europe might mean 100, 120 or 144 depended on context. This holds true even when a hundred is written in ciphers (100).

Another important aspect of historical data is 'fuzziness'. Data are fuzzy, when seemingly accurate data are in fact the expressions of an underlying distribution. An example could be that a document gives the age of the persons mentioned in multiple of five years. In a document like that the age 15 would probably mean something like an age between 13 and 17, but to determine or approximate the actual distribution of the ages underlying the number 15 would be a research project in itself.

The software needed to handle context sensitivity and fuzziness is clearly something quite different from commercial database management systems.

The historical workstation

One theoretical attempt at approaching the uniqueness of historical computing is the historical workstation. The technical definition of this term is:

A desktop computer, which
has access to data base management software, which is able to administrate very different structures of information, allowing to put into a data base arbitrarily large collections of sources, keeping as much information of the original and applying as little coding, as is economically feasible for the project producing the data base,
has access to a set of data bases, which contain background knowledge specific to historical research,
has access to a large number of read only data bases, being equivalent to traditional printed editions of source material,
contains sufficient Artificial Intelligence subsystems to make the interaction between the forementioned capabilities transparent to the user,
has a very highly integrated interface between the data base management system mentioned and a desk top publishing system and

[6]The first quarter of the seventeenth century. Information on the effect of the Kipper Wipper time in Denmark can be found in: H.J. Marker: Sletdalerbegrebet i første fjerdedel af 17. århundrede, Historie XV, 4, 1985, p. 633 et seq.

[7]The example is given in Manfred Thaller: Why do we need a Theory of Historical Computing?, in Charles Harvey (ed.): Proceedings of the second annual conference of the Association for History and Computing, Manchester 1988.

a similar interface to statistical software.[8].

This definition shows that the historical workstation is a software concept more than a hardware concept. Although a piece of hardware is necessary and will have to meet a number of minimum requirements, the software for the workstation is likely to be made available on several different brands of hardware. Most of all, the historical workstation project can be seen as a framework for the co–ordination of international co–operation on software development in the field of history.

The project concerning the realization of a historical workstation was generated at the Max Planck Institut für Geschichte in Göttingen, Germany and it is progressing with the co–operation of a great number of institutions all over Europe. In Göttingen they have also managed to get funding for some of the major building bricks of the project from Volkswagen Stiftung and IBM, Germany.

The historical workstation project serves as a common standard for program development. The condition being that the bits and pieces should fit into the entity in the end. Another important prerequisite of this project is the exchange of data, preferably over national boundaries.

Beyond the Historical Workstation

In my view the perspectives arising from the historical workstation go far beyond tool making for historical research. I think that there is a reasonable chance that the tool will change the trade. This would happen when results of research projects would be expressed in software running on a commonly available workstation. A procedure like that would be especially relevant when the project in question were dealing with structures, relationships, values, models or other things which are easily expressed on a computer. In this way the next historian would be starting where the former ended instead of going through the same research process again. Or in other word by using software to express research results, history would acquire the equivalent of the symbolic language used by mathematics, physics and the other natural sciences.

A trivial example would be: When somebody expressed the dependency of exchange rates between coins of a specific territory on time and context in a piece of software for the workstation, future research would have two options, either use the software as a given building brick in further economical history studies, or refine, correct or expand the software by means of intensified studies in the history of currencies.

The example is deliberately very trivial but similar examples could be made where the original research result was in prosopography or even the structure of society.

This extension of the historical workstation concept is my own and should not be held against any of the other participants of the project. I also confess that the whole concept of the equivalent of a symbolic language for history is heresy[9] as opposed to the common scientific standard of history today, where a historian is expected to understand every detail of the picture, which he is presenting[10].

I would also like to clarify that I do not think that the way of making historical research which I have outlined above should be considered the only "right" way of doing

[8]This version of the definition is quoted from Manfred Thaller: Data Bases v. Critical Editions, p.5

[9]The word 'heresy' was given to me by Deian Hopkin.

[10]I cannot resist the temptation to add: "rather than to present a very large or interesting picture." (Sorry about that.)

research in history in the future. For a number of research questions it would not be possible, for a great number of others it would not be the most practical way of working. What I do think however is that this way of doing historical research would give us the possibility of getting answers to some questions, which we can't answer today, and even some which we don't even get far enough to ask by present day methods.

The Role of the Data Archives.

The data archives of north western Europe are national social science data archives. They have been in existence for a couple of decades now, and are generally preoccupied with the conservation of social science survey type data. There are some differences between their operation, but the similarities are perhaps greater.

The traditional role of the data archives, to document and preserve survey type data are naturally of great potential interest for historians. The survey data are unique historic sources, as they are dealing with the attitudes of the general populace and are using scientifically well defined methods to analyze them. This role of the data archives will be of increasing importance as the data materials archived in them become older.

Data materials from quantitative history usually fits in more or less easily in the general pattern of data archiving. As they are numeric they can usually be forced into the traditional data description. Some historians clearly feel that the study description schema is missing the point about what their study is actually about[11] and finds the methods of traditional data processing at the data archives a bit strange. Anyhow some of them are quite willing to live with that, and a few even finds the data archives useful.

When it comes to data materials with complex data structures, the traditional data processing methodology at the data archives runs into difficulties as it is designed to cope with survey type flat file data. I don't think that the data archives have found a common way of handling of material with complex data structures yet. One possibility is to store the material and redistribute it on an as is basis, which is quite feasible but from the historians' point of view it must be hard to understand, that the more complex and often more valuable materials recieve a much less sophisticated treatment than the simple and perhaps not all that interesting materials.

Historians also generate a number of nonnumeric data materials. There is some uncertainty about whether they are within the scope of the data archives, but in many cases the fruitful historical research will be carried out with a combination of numerical and nonnumerical material. For historians working in that way it would clearly be most practical to have all their material archived at the same place. Further the dividing line between text and other complex data structures is becoming increasingly blurred with the advent of retrieval systems, where an entire text is stored and parts of it is marked up as fields for retrieval.

[11]Some discussions have been going on at three annual conferences on the exchange of data within the historical sciences. The outcome of these discussions is a proposal for items that ought to be included in a description of a historical data material. One version of this proposal can be found in Herbert Reinke, Kevin Schurer & H.J. Marker: Information Requirements and Data Description in Historical Social Research. A Proposal, Historical Social Research 42/43, Köln 1987

In the ideal world all the problems of the historians could be solved by the national data archive. In some of the fields the expertise is already at hand, in some of the others it would have to be acquired, but still no other institutions have a better foundation for serving the needs of the history and computing people. The national social science data archives are centres for the exchange of data between researchers and they have for years now been operating an international data exchange network exactly like the one discussed by the historians today.

Unfortunately we are not living in the ideal world, and in the world we live in one of the problems is funding. In some European countries it is regarded as unlikely that the increased load of work which will result from increased interest of the data archives in the field of history will be compensated by an increase of staff and funding.

In Germany a solution is on the way as the Zentrum für historisches Sozialforschung has been reerected as a department of the Zentral Archiv in Köln. The traditional interest of the Zentrum is in the field of quantitatively based social and political history, and as such they fit in very nicely in the standard procedures of a social science data archive. By the way this also happens to be one of the fields of history and computing which have the strongest traditions.

In the Netherlands the possibilities for making a historical data archive with close connections to the Steinmetz archive are being investigated, and as I understand it an experimental historical data archive is being made as part of these enquiries.

In Norway they have NAVF's EDB Senter for Humanistisk Forskning [12] this institution is not devoted to history alone but to the humanities as a whole. They even feel that their major responsibilites lie in the other fields of the humanities because as they put it: "the historians are already fairly advanced." Anyway a number of their projects are in the field of history. The EDB Senter is placed the same block as the Norwegian Social Science Data Archive (NSD) in Bergen.

Beside the mentioned a number of other European institutions like the large Swedish demographic databases, the Norwegian central for the registration of historical demographic data in Tromsø, the Cambridge Group for the History of Population and Social Change, Cambridge, England, and the Max Planck Institut für Geschichte in Göttingen, Germany, serve as centres for and to different extents also as data archives for computing in history.

[12]Norwegian general research council's EDP Centre for research within the humanities

The DDA and History

The involvement of the DDA in the field of history goes back far. One of the reasons for the establishment of the DDA was that the social science surveys were of considerable potential historical interest. Some of the earlier DDA studies have perhaps reached an age where their prime interest may lie in the field of history. DDA studies are numbered in chronological order with DDA-0001: Danish Omnibus Survey 1982, being the first study registered at the DDA. This happened in 1982. The oldest study at the DDA of merely historical content is DDA-0018: Danish Politicians: Members of Parliament 1849-1968. The oldest purely historical study at the DDA is DDA-0034: Reconstitution of Biological Families. Sejerø 1663-1813. This study was registered in the DDA in 1976. Of considerable importance for the DDA's interest in history was the contact with Hans Christian Johansen.

This contact dates back to the very start of the archive in 1973[13].

A number of historical research projects have been carried out at the DDA or with assistance from the DDA. Among the DDA generated projects are Prices and Wages in Eastern Jutland 1571–1661[14], which is a data collection project on prices found in official ledgers given from local to central administration, and Bilantz 1660, which is a registration of the creditors of the crown in the year 1660[15].

As regards projects in which the DDA has participated the following three amongst others deserve mentioning: Population History of Greenland, 1800–1930, (DDA–0235), Court Records: Elsinore 1612–1730 and Falster 1665–1718, (DDA–506) and Koldinghus Fief Ledgers, (DDA–755 and 830). **DDA Studies in the Field of History**

The 1986 DDA catalogue lists 70 studies in the category history and demography. The holdings have increased somewhat since then, and as mentioned above the older social science studies are aging.

The studies cover very different subjects and research methods. Among the more well–known of our studies are the ones produced by Hans Christian Johansen: DDA–0038: Sound Traffic, 1784–1795, based on the customs register from the Sound, DDA–0106: Reconstituted Families in Selected Rural Parishes, 1741–1801, DDA–0230: Urban Population of 18th Century Denmark, and DDA–0778: Danish Economic Statistics 1814–1980. **International Co–operation**

Besides being a place for the storing of machine readable material the DDA engages itself in national and international cooperation on software development in the field of history, standardization, computer usage in research etc.

The DDA has followed the workshop series on standardization and documentation known as the Thaller group from the beginning in Göttingen, 1985, over Graz, 1986, Paris, 1987 and we will also be present in Köln 1988. DDA has a member in a subgroup of this line of conferences working on the standardization of data description in the field of history.

The DDA also has one of the two Danish members of the Nordic Demographic Data base Working Group, which is a group of six from Denmark, Norway and Sweden, and is responsible for the organization of a triannual Nordic conference on computing in history, and DDA has a member of council in the Association for History and Computing. Besides the DDA is in close contact with the Historical Workstation developers group.

The Historians and the Data Archives

It is not entirely clear which role the historians would like the data archives to play. Although some organizing is taking place among the historians within the framework of international bodies like the Association of History and Computing, the historians are clearly not speaking with one voice as yet.

The quantitative historians are usually quite happy with archiving their data materials at a social science data archive, while the historians of nonnumeric persuasion are less convinced that the data archive staff possesses the expertise needed to deal with their data. Unfortunately there is quite some animosity

[13]Rapport fra Referencegruppemode 1, April 1973.

[14]DDA–1066. Further information on this project can be found in H.J. Marker: Danish Prices and Wages and the Micro Computer, in Deian Hopkin and Peter Denley (ed): History and Computing, Manchester 1987, pp. 89 – 95

[15]Per Nielsen: Beretning for DDA i finansåret 1985, DDA–NYT no 36, 1986, p. 25

between the historians of the quantitative and the nonnumeric approach in some countries.

The diversity among the historians means that in some countries it will not be without difficulties to establish the social science data archive with its clearly quantitative tradition as a generally accepted depository for historical research materials.

The Future?

It seems clear to me, that the national social data archives and the funding authorities behind them have to take a stand soon on whether or not they would like to see a network of institutions relatively similar to but separated from the social science data archives dealing with historical data or they would rather have general data archives for social science and history[16].

In my view a great number of historians would be better served by having the expertise and the data needed for their purposes at the same place, rather than having some of it at the historical data archive and some of it at the social science data archive. Some data materials are of interest to historians and social scientists alike, and it would be most practical for the researcher wanting information on a specific data material to know that this expertise was to be found within one institution rather than in one institution for this data material and in another for that.

Also an institution dealing with history as well as social science would have to be larger than institutions dealing with only one of the sciences. This would be of advantage in most countries as it would create possibilities for a better scientific setting. My personal hope would be that the enlarged national data archives dealing with social science and history alike, would act as power houses for developments like historical informatics.

If a solution with separate data archives for history and social science is preferred, a number of routine tasks and a great deal of development of methodology and software tools would be exactly alike in the two data archives. Coordination of the efforts would result in savings in time and effort. The same would probably hold true for administration. The possibilities for saving time and money should appeal to the funding authorities.

Whether or not the social science data archives should want to be increased with a historical branch is another matter. To some extent it is a matter of personal preferences: does one prefer a wider range of activities over larger degree of homogeniety.□

[16]In some countries this problem does not exist in the literal sense as history is regarded one of the social sciences, while in other countries history belongs to the humanities. Anyway the problems do exist when it comes to historical data materials which are clearly out of line with the material usually archived at a social science data archive.