A source-oriented approach to history and computing: the relational database

Greenstein, Daniel I.

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Historical Social Research, Vol. 14 — 1989 — No. 3, 9-16

A Source-Oriented Approach to History and Computing: The Relational Database

*Daniel. I. Greenstein**

The computer's place in historical research has yet to be firmly established despite 30 years of its use by historians. This is largely due to the fact that the computer is viewed as the naturally of so-called »new« historians whose use of social science models and quantitative methods has challenged »traditional«, narrative approaches to history and engendered bitter controversy over the fundamental nature of clio's craft.(1) In this debate, the computer is painted by the traditionalist with the same brush as the renegades, and revered by the renegades as a weapon in their challenge to the traditionalists. It is not often treated by historians for what it essentially is: a tool which, like many other tools, has some general utility in the study of history. In his much quoted presidential address to the American Historical Association in 1963, Carl Bridenbaugh reacted strongly against the pioneering use of computers in history. It was not, however, the computer »per se« to which he objected, but »the Bitch-Goddess QUANTIFICATION«.(2) Over the course of the next decade, Bridenbaugh and traditional historians like him were increasingly on the defensive against a groundswell of new history. By 1973, Le Roy Ladurie could comfortably forecast that historians in the 1980s »would have to be able to programme a computer in order to survive«.(3) Once again, however, it was not the computer that was at issue but the approach to history that was facilitated by its use.

By the mid-1970s, the new history was quite clearly in decline; many of its more eminent practitioners began sounding a retreat to narrative or admitting that borrowed social science models and methods can only ever supplement never replace the total immersion in the archives that traditionalists have always advocated as a defining principle of their discipline. Now, with the move away from the new history in full swing, literature on historical computation is increasingly confined to journals which are prepared for and consumed by a small cadre of the converted. To all but the cognoscenti, its language and meaning is impenetrable. The computer is

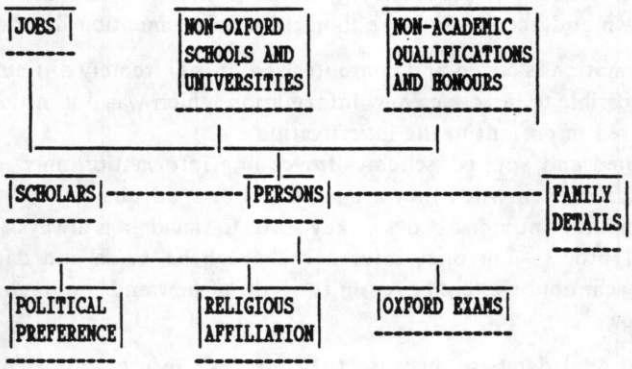
* Address all communications to Daniel I. Greenstein, Corpus Christi College, Oxford, OX1 4JF, Great Britain.

thus in danger of being associated by historians with this particular rift within their profession and, more narrowly, with the sophisticated and, for many, unapproachable methods of quantitative history. So long as this is the case its role in historical research will neither be firmly established nor widely accepted.

To counter this trend towards what can only be described as marginalization, the computer must be shown to be useful in what I shall term a »source-oriented« approach to history. This claim is based on the following premises. Firstly, historians do not on the whole specialize in the two primary characteristics of computation: technique and method. Instead, they specialize in the use and interpretation of historical sources. Secondly, historians will handle the same sources differently according to their particular research interests, and to their explicit but more often implicit theoretical perspectives on the nature of history. Thirdly, any one historian's interpretation and use of a particular source may vary over time according to intuition and to the discovery of new evidence. In short, historical research sees the historian engrossed by the archives, engaging the sources therein in a dialogue which is highly personal. History, then, is in some large measure, subjective despite some radical but no longer fashionable claims to the contrary. The computer, if it is ever to be seen as having some general utility for the profession must replicate, indeed facilitate the dialectical relationship (in a Socratic not a Marxist sense) that exists between the historian and his sources. If, on the other hand, the computer continues to be dressed up and sold to historians in terms of its application of rigorous methods and standards to the use and interpretation of historical sources, it simply will not wash with the larger body of the profession.

Relational database technology goes some way in providing a more source-oriented approach to historical computation. The scope of this essay prohibits a comprehensive account of what a relational database is and how it works. Suffice it here to say that a relational database is seen by its users as a collection of tables. (4) The diagram in Figure 1 shows the tables of a relational database being used in a study funded by the Leverhulme Foundation of the social origins and career destinations of a sample consisting of 15,000 of Oxford University's twentieth-century students. The central PERSON table consists of the names and vital statistics (date and place of birth and death) of each individual for whom biographical information was collected. Other tables include information on educational attainment, non-academic qualifications, jobs held, political and religious affiliations, and so on. Biographical information in these tables is joined to the relevant individuals whose names are listed in the PERSON table through intermediary tables (not shown in the diagram) which also contain references to the source from which such information was derived.

Figure 1: **lough diagrai (excluding intermediary join tables) of database wdel; study of Oxford University's twentieth-century students**



There are four principal advantages that the relational database offers the historian. Firstly, they permit historical sources and the aims of historical research to determine computational method and technique. This is something of a departure in and of itself insofar as computer-assisted research is so often shaped by the rigid constraints imposed by available computer software. Take, for example, hierarchical database software which insists that data be represented in one table only. That table consists of one row for each record in the database and as many columns as there are variables that are likely to be considered in a particular study. The problem, of course, is the number of variables involved in any study arc determined through the historian's familiarity with the historical source material and so is constantly in flux as more information is uncovered or as new perspectives are adopted. With the relational database, this »dialec-tical« interpretive process is better catered for. There is no limit, for exam-pie, to the number of records concerning individuals' jobs that may be added to the jobs table represented in Figure 1 above. Further, there is no reason that an entirely new category of information, say on forms of edu-cational funding, could not be added to the database simply through the creation of a additional tables which are linked to the PERSON table.

A second advantage that the relational database offers the historian is that it allows for historical information to be stored and analysed in its raw (read textual) form. A glimpse of a small portion of the jobs table is pro-vided in Table 1. There, it may be seen that occupational information is presented precisely as it was discovered at source. Thus, the relational

database establishes something akin to an archive in which the source is available to any number of different interpretations and research designs. Compare this with more rigidly structured database software which force the historian to sweat down his data into a series of standardized keywords or even numeric codes. The problems with coding historical data are already well known and need little more than the briefest mention here:

> once information is coded and computerized, it is extremely difficult if not impossible to introduce new information which was not initially considered important to the investigation;
> once adopted and applied, schemes for coding information may rarely be altered, even when they are shown to be inadequate;
> the assignment of numeric codes or keywords to raw data is always an inferential process, but once inferences have been made and data coded they cannot be altered even in the light of new and convincing information.

Thirdly, relational database, because they preserve information in its largely unaltered form are themselves something of an archive whose material is accessible for secondary and comparative analysis by any number of different users. Each is able to bring to the data his or her own research interests and expertise. Were the data in the study to be presented to them in a highly structured form (whether keywords or codes) they would be denied that one fundamental element which comprises so much historical research. That is, the chance to engage the raw material, to ask their own questions of and derive their expertly informed interpretations from it.

It has often been said that the vast majority of computer-aided historical research produces results which rely on the simplest quantitative techniques: counts, percents and averages. Such procedures do require that historical data be grouped together in some meaningful way. But, as described above, the process of categorizing historical data in non-relational databases is both controversial and intractable. The fourth advantage of the relational database, then, is that it enables data to be categorized for quantitative analysis but in a way which docs not alter the raw data.

Together, Tables 1-4 provide only the end product of one example of data normalization as it was carried on the database whose overall design is shown above in Figure 1. The interpretive question at issue was the extent to which college and university scholarships were used at Oxford to help poorer students with proven academic merit to achieve a university education and how, if at all, this changed over time. The PERSON table lists the names of a sample of Oxford University's twentieth-century students. Since the study in question was interested in university education and inter-generational mobility, it also includes the names of those students' fathers and spouses. The SCHOLAR table, on the other hand, pro-

Historical Social Research, Vol. 14 — 1989 — No. 3, 9-16

Table 1: JOBS

job code	job title	firm name	department	place
1	Accountant	IBM	planning dept	Basingstoke
2	Editor in charge	The Times	foreign corr.	London
3	Professor	Harvard	History	Cambridge, Mass
4	Research Chemist	ICI	paints	
...				

Table 2: PERSONS

name code	surname	first	second	third
1	Plates	Clifford	Lowe	
2	Stones	Robert	Henry	Cyril
3	Barnes	John	Arthur	
...				

Table 3: SCHOLARS

schol-code	scholarship
01	C Plummer Exhibitioner
02	Classics Scholar
03	College Exhibitioner
04	G C W Winter Warr Scholar

Table 4: normalized analysis of scholars and coroners' fathers'occupations Corpus Christi College, Oxford, 1880-1974 (1=2,987)

	1880-1913		1914-39		1940-74	
OCCUPATION	scholar	commoner	scholar	commoner	scholar	commoner
gentry/any	11t	20*	St	lot	6t	7t
clergy	26	17	13	9	4	3
civil service	5	5	9	9	12	9
lav	4	14	7	10	3	6
medicine	3	4	5	6	4	5
teaching	8	5	9	14	14	10
other professional	8	3	8	7	10	7
all professional	23	26	29	37	31	28
finance	2	4	5	5	4	4
commerce	13	10	10	13	10	12
industry	5	6	8	7	JL	10
all business	20	20	23	25	26	26
clerical/shopkeeper & working class	8	1	16	5	17	16
not known	7	11	2	5	4	11
TOTAL PERCENT	100	100	100	100	100	100
NUMBER OF CASES	257	410	233	370	584	1133

vides a list of various scholarships and prizes that were available to help students pay for their university education. Finally, the JOBS table gives a complete list of all the jobs known to have been held by the university students in the sample as well as those held by students' fathers and spouses.

Owing to limitations of space, Table 3 does not show the two intermediary tables which link the people in the PERSON table with their scholarships and jobs held in the SCHOLAR and JOBS table respectively. Nor does it show the supplementary table that was created in order to provide for the requisite categorization of occupations. Very briefly, this supplementary table consists of two columns was created in the database. The first column of this table is entered up by the software itself which inserts into it a concatenated version of the data which appear in the substantive columns of the JOBS table - job title and firm. The second column contains a code indicating that the job in question was in one or another of several job categories, and this must be updated by the historian conducting the analysis. From this point forward, it was only necessary to follow standard procedures for so-called »data normalization« in order to produce the end product: a table comparing scholarship holders with non-scholarship holders by their fathers' occupations for the periods, 1880-1913, 1914-1939, and 1940-74.(5)

Three points require emphasis. Firstly, the analysis took place without altering the raw data. Secondly, repeating the analysis but basing it on entirely different criteria for categorizing occupational information requires simply that the contents of the second column in the supplementary job-coding table be altered. Thirdly, the figures in the result table, expressed as either percentages or counts, could easily be loaded in batch into any number of statistical tables where they could be multiplied, added, regressed or tested for variance.

Despite the obvious utility of this procedure, it is difficult to advocate the view of historical research upon which it is based. That is that computer-assisted historical research comprises a singular and relatively brief strategic raid on some archive where as many data as possible are captured, and neutralized in a fixed and unalterable pattern which is conducive to some foregoing analytical procedure.

The relational model, it seems replicates a more traditional approach to history; one in which the process by which the historian gets to know his sources, and that by which he analyzes them are worked out together and in relation to one another. One cannot, therefore, advocate the relational database as a means of storing historical data in its textual form as a prelude to its being rammed into a flatfile. Rather, as a kind of archive in and of itself. One, which like traditional archives, may be continually updated, and freely and selectively explored by historians.

Notes

1. R.W. Fogel approximates relations between traditional and new historians to »cultural warfare« in R.W. Fogel and G.R. Elton, Which Road to the Past: Two Views of History, (London, Yale University Press, 1983).
2. Carl Bridenbaugh, »The Great Mutation«, American Historical Review, 68(1963).
3. Emmanuel Le Roy Ladurie, The Territory of the Historian,(Chicago, University of Chicago Press, 1979).
4. E.F. Codd, »Relational Database: A Practical Foundation for Productivity^ Communications of the ACM, 25:2(1982); E.F. Codd, »Is your DBMS Really Relatoinal?«, Computerworld, (Oct 14, 1985); CJ. Date, Relational Database: Selected Writings, (Addison-Wesley, 1986).
5. Normalization - the procedures through which relationally structured data are transformed to a hierarchical structure - has received sufficient treatment elsewhere by Codd and others to mitigate against reproducing their work here.