Preserving Historical Databases and Facing Technical Change: Common Issues for Social Historians and Archivists

by **GORDON DARROCH** and **SUE GAVREL***

Introduction

This paper addresses two sets of issues that arise for historical analysis as a consequence of the rapid development of computer technology. In the first place, we are concerned with access to computerized files of historical data for the purposes of secondary analysis. We make a number of specific recommendations regarding documentation practices for individual projects that would enhance preservation of and accessibility to historical data files in the future.[1] Secondly, we discuss the challenges raised for archiving by the recent technological changes in the creation of electronic records. Though these can be treated as separate issues, they are united by the fact that both arise from the rapidly evolving computing environment of historical research and archiving, and they raise common issues regarding the conservation of historical data. The first should engage social historians and the active interest of archivists, especially those responsible for preserving non-government records. The second and more general issue now commands the attention of archivists, and should be of vital interest to social historians.

Archiving in A Microcomputer Environment

The growth in the hard disk memory capacity of microcomputers and the development of powerful and well-documented database software has greatly enhanced the ability of individual scholars to create self-contained, computerized files of historical data.[2] Coupled with the now twenty-year-long extension of historical projects based on the systematic collection of nominal historical data, these developments have fostered a significant "cottage industry" in the production of historical databases in the hands of individual scholars. We are not able to guess the numbers of such data files that have been created and maintained by individual investigators, either loosely tied or wholly unconnected to an institutional computer centre or other data archive. There are sufficient numbers in Canada alone, however, to warrant concern about adequate documentation and the implications for general, scholarly access in the future. Certainly, given the rapid growth of microcomputing, there is every reason to suppose that this research "cottage industry" has international dimensions and is expanding.

 Archivaria 34 (Summer 1992)

In contrast to these individually-conducted projects, there are the well-documented databases of nominal historical data generated by relatively large, team-conducted and publicly-funded projects. There are, for example, the United States historical census samples, the Umeå demographic database of Sweden, the Casalecchio database of Italy and the immense database of the Saguenay region of Quebec, which is in the course of being extended to the entire province.[3] These projects provide exemplary cases of historical databases which are both fully documented and made fully accessible to the intellectual community through established procedures. Other examples of accessible historical data files are those which, following the precedent of publicly-funded survey research, have been deposited in a national dissemination centre, such as the Inter-University Consortium for Political and Social Research at the University of Michigan, the Social Science Research Council Archive at the University of Essex, or York University's Institute for Social Research. The less ambitious and costly projects, however, especially those fostered by microcomputing, have not been routinely deposited and documented, nor are there, to our knowledge, national inventories of such files.

The proliferation of smaller-scale historical data files raises the question of their status as archival records available for further analysis by the scholarly community. Despite the fact that archives and archiving practices are integral elements of history as a discipline, access to individual historical data files for secondary analysis has not yet become a common expectation. Perhaps the strong emphasis in the discipline on originality and individual scholarship still tends to conspire against the practice, but there are many reasons to encourage wider use.

First, increasing numbers of quantitative social historical projects represent a potentially rich, accumulating resource for secondary analysis. Secondly, the data collected in an initial project are often subjected to a relatively narrowly focused and, hence, limited analysis by the principal investigators. Thus, archived data provide opportunities for analysis that were not originally recognized, employing methodologies and theoretical perspectives not initially contemplated. Thirdly, even if one must ultimately create a unique database, some initial analysis may be undertaken using existing files, in order to refine conceptualization and research design prior to the new data collection. Fourthly, and specifically in the absence of national directories of existing historical data files, there is the increased prospect of duplication in data collection from a given source. Fifthly, that such files may remain unknown, except through the author's publications, raises the issue of the unnecessary, perhaps unintended, privatization of a potentially public academic resource. Our concern with the limitations on wider use of data collected by individual researchers does not gainsay the latter's priority in analysis and continuing interest in the data. Finally, the chances of losing unique historical files are greatly enhanced by the absence of routine archival procedures practised by individual researchers. One may be aware of files that have been rendered useless or seriously limited as a result of simple computing error, failure to provide backup files, failure to update storage media or simply through inadequate documentation.

There seem to be three options for maintaining and disseminating historical databases. Two of these have already been noted and apply to the larger, well-known and publicly-funded projects. In the first place, some may have sufficient support to ensure adequate documentation and dissemination, deposit of data in an environmentally-controlled computing facility and routine attention to keeping pace with technical improvements in

electronic storage media. Normally, these would be university computer centres or their equivalent, which have institutional funding and personnel.

The second option is to deposit data in a well-established survey/research centre, such as those at the Universities of Michigan and Essex or York University. This is undoubtedly the preferred option, since these centres can provide standardized documentation, they have the appropriate physical storage and tape and disk management facilities, and they provide a greater prospect for secondary analysis than any alternative arrangement.[4] Neither of these options, however, can be expected to deal fully with the growing numbers of smaller-scale historical databases constructed by individual researchers.

A third option is worth proposing in the absence of institutionalized incentives among historical researchers to deposit computerized data files in such archives. This is to create national directories of historical databases, leaving the maintenance, documentation and dissemination of the files in the hands of the creators or their institutions. Clearly, this option makes some sense in terms of the specialized nature of the databases and the still modest, though probably increasing, interest among secondary users over the next decade or two. The inventory itself would presumably enhance the possibilities for secondary analysis. The newly-formed Committee on Computing of the Canadian Historical Association (CHA) has recently discussed the recommendation that an existing, experimental union list of social science data in electronic form be systematically extended to historical projects. The compilation of the union list was initially supported by the National Archives of Canada, but is intended to be supported beyond the pilot stage by subscription fees. We can only urge the CHA committee to pursue the initiative and to seek the financial support it would require.

We shall be addressing the question of minimum adequate standards of documentation that would be required for such separately held, but centrally indexed files. Here we note that the success of such a inventory rests on an initial, systematic survey of the community of principal investigators who may have created, or are in the process of creating, historical data files. Such a survey would probably best begin with a relatively small number of selected, knowledgeable informants in the academic, data archive and academic funding communities, who can produce a list of principal investigators known or believed to have created historical data files in electronic format. The latter would then be surveyed in order to attain the basic information on the nature of the files, their documentation and current accessibility.[5] Such a survey cannot be treated merely as a casual inquiry, if it is to serve as the foundation for a continuing, routinely updated inventory of holdings. Social survey experience teaches that successful surveys require careful planning, attention to detail and financial support for both initial and subsequent contacts.

Documentation Standards and Practices

Whether national inventories or central repositories for historical databases can be established or not, the archival value of such files depends on their documentation. Though such a claim is obvious, the fact is that individual historians constructing electronic data files have not always met minimal standards of adequate documentation. Some significant historical files are currently quite impossible to use given the limitations of their documentation. We offer recommendations on the basic practices of data collection and on minimal documentation standards. It is useful here to distinguish between primary and secondary documentation.

Primary Documentation

The problem of adequate documentation for relatively small, individually-conducted historical projects is simply that it is the responsibility of the principal investigator—along with data collection, analysis and publication of results. In addition to the relative lack of priority that will normally be accorded these seemingly clerical tasks by researchers, a number of factors tend to conspire against full file documentation. One factor is simply the lack of widely shared knowledge of archiving practices for electronic records. Another is how surprisingly simple it is for principal investigators to lose track of essential aspects of their own documentation trail, for example, of precise definitions of missing variables or the details of recoding variables. Furthermore, few researchers make full provision for the costs of and time required for documentation in research grant proposals, much less consider the continuing costs of maintaining and occasionally disseminating databases over the long term.

Though the specific aspects of each case are obviously unique, some general recommendations for the maintenance of primary documentation of historical databases may be offered. At the outset we urge that whatever the nature of the source documents, normal practice should entail the *complete transcription* of original data. The increased capacity and speed of computing, including microcomputing, largely renders irrelevant the questions of memory capacity. Those who collect data from historical sources can seldom imagine the full range of possible analyses to which the data may eventually be subject. Coding decisions always entail loss of information, and are better left to a phase in the analysis of the database.

Secondly, as part of the standard code-book, giving definitions of all variables in the file, provision should be made for maintaining a detailed diary or "audit trail" of the file construction, and of data-cleaning and error-checking procedures. The "audit trail" serves as a log tracing variable redefinitions, case deletions and additions, and decisions on missing data. Especially in a microcomputer environment, there is a temptation to rely on the apparent clarity of the decisions at the moment they are made on the screen, rather than separately recording them. Finally, some investigators make it a practice to maintain computer listings as a log of variable and file changes. Though useful for the recovery of file changes, listings alone are wholly inadequate for deposit as file documentation, since they place impossible demands on secondary users.

Secondary Documentation

Basic file documentation is considerably enhanced by the provision of simple technical documents about the study design and data collection. The main secondary document should be a relatively detailed report on the data collection procedure, which provides an interpretive context for further analysis of the data. The report should identify the *character* of the data and its provenance, not just the specific aims of the original project. This is particularly important, since many studies amass a wider range of data than is required by the original project. When one is creating an electronic record of original manuscript sources, the "marginal cost" of transcribing variables beyond those of immediate interest is modest; one is very unlikely to return to a source document in order to add some missing evidence. The report should include at least the following elements: the dates of the project; the definition of the source data; if a sample was employed, the specific aspects of its design and of the sampled population; and a basic

description of the method of data collection. In cases where the data have been sampled, a *full* description of the sample design is essential, including reference to any forms of stratification and clustering, as well as providing definitions of sample eligibility and population and sample exclusions. Reference to the accuracy of estimates as compared to known population parameters is essential. Exact specifications of "weighting" variables are required where cases are to be adjusted in order to accommodate over- or under-representation of specific groups in the population.[6]

We also advise reporting facsimiles of the formats for computer screens that were employed in the data collection process. The report would be further enhanced by a set of marginal tabulations for each variable. These tabulations are a particular help to potential secondary users in determining the possibilities of analysis. Finally, assuming dissemination of the files on magnetic tape, the tape specifications and file structure of the data must be clearly reported (and recorded on the tape itself).

Concerning the file structure, we urge that the data be maintained or deposited in a "flat" format, that is, as raw data rather than as a file dependent on any particular computer software. There are simply too many continuing alterations in computer software to ensure that data can be "read" in the future, if it is not in a raw form. Despite the relative recency of archives of electronic records, there are already instances where whole files have been rendered useless by changes in software. This central question of the implications of increasingly rapid technological change is taken up in the following section.

As an addendum to a file, it is useful to include citation to publications based on the data. The technical report can also serve the principal investigator by specifying clearly any restrictions that are to be placed on the dissemination of the data or on analysis.

Facing the Future: Source Data, Archiving and New Information Technology

Trends in Information Technology

The production of source records has been fundamentally altered in the last ten years by the rapid penetration of microcomputers into the market and, in particular, into the office environment in both the government and corporate private sectors. In this section we turn to the challenges posed by these changes for the conservation and archiving of source documents, that is, the question of what kinds of "records" historians and archivists will face in the immediate future. There are also pressing questions about the possibilities of conservation, access to and analysis of electronic records. The questions posed here must engage both archivists and historians.

It is now commonplace to note that managers, officers, and clerks often have as much computing power on their desks today as the mainframe computers provided in the 1970s. Accordingly, the ability to manipulate, access and disseminate data is in the process of being massively decentralized towards the records creators. Databases, images, and text are routinely created, stored and redisseminated by those who would not presume to call themselves electronic data processing specialists. Moreover, the character of records themselves is being transformed in the process.

Linked to the organizational penetration of microcomputers is the development of Local and Wide Area Networks, which further encourage on-line creation and revision of documents, although only the final version is available in either hard copy or electronic

format. Voice annotation of such documents is now possible and, no doubt, will expand in use, especially in managerial and policy-making settings. In addition, there is a growing capability for accessing on-line database systems, transferring specific data from one source or document to another, or combining data from a variety of sources.

Thus, the ability to document precisely the evolution of policy and its changes through administrative channels is now seriously impaired by the variety of sources used in the process, by the lack of reference to specific sources, and often by the loss of any electronic record of key changes in the emerging document.[7]

Furthermore, the most recent developments in software, such as fourth-generation languages and so-called expert systems, which facilitate complex inferential and problem-solving procedures, will have a significant effect on record-keeping in some organizational settings, for example, in assisting with complex planning and scheduling tasks, or with the diagnosis of disease. Clearly, such developments complicate our definition of a source document and, indeed, tend to make invisible otherwise routine documentation of organizational functioning.

The extent of the routine use and integration of electronic databases and office microcomputers will be enhanced by the developments in data storage and retrieval capacities, and by the emergence of more widely adopted communication standards. With respect to the first, currently available optical disk systems greatly increase data storage capabilities. Optical disks are readily stored in the office environment, thus obviating the need for more controlled, separate storage facilities, and can serve to link very large databases with office microcomputer systems.

With respect to communication standards, the lack of compatibility among hardware systems still represents a major barrier to the routine use of integrated databases. There is, however, significant movement towards the development of international standards, such as adopting the standard protocols of the so-called Open Systems Interconnection for relating systems with different hardware structures. Moreover, telecommunications companies and hardware and software vendors are moving away from proprietary standards for their products. The development in software of a standardized document architecture and interchange format provides for the exchange and processing of documents by many users. Other standard protocols are being developed for databases, under titles such as the "Information Resource Dictionary Standard" and the "Standard Query Language."[8] These developments will further promote integration of computing and database systems and yet further complicate the notion of a source record for archival purposes. We also venture to argue, however, that the increasing agreement on communication standards will be the key to enhancing the archiving effort in this emerging computer environment. This issue is addressed below.

In general, the increasingly standardized communications among computer systems presents a challenge to archivists of electronic records. The cost of collecting, storing and retrieving data from several sources is rapidly diminishing. The use of multiple sources of data is reflected in the trend towards integrated systems and applications that can serve a variety of analytic and administrative purposes.[9] Each development clearly requires a redefinition of the traditional archival functions of identifying and tracing "source documents" for purposes of appraisal, documentation or dissemination. We argue that if there are to be "records" available for analysis in the future, archivists acquiring electronic records must routinely have the opportunities and resources

available to study and trace the implications of these changes as an integral aspect of their professional practice.

New Problems in the Preservation of Electronic Records

The electronic records of the 1970s differ very significantly from those created in the 1980s and from those that will be produced in the current decade. The first problems faced by archivists are simply the volume and complexity of the electronic data that must be considered. This is particularly the case for government and private sector corporate records archivists.

The sheer volume of electronic data is increasing rapidly and is generated within all forms of administration and management. Regardless of the increased storage capacities of various media, we may simply be faced with the prospect of being unable to conserve data that would normally be appraised as having archival value. In addition to the volume, the complexity of the information systems being created, particularly in the scientific field, may also prove to be a serious stumbling-block for traditional institutional and national archives. These considerations alone suggest that we need to reconsider traditional archival practices and the roles of both archivists and historians in the emerging electronic environment.

There are other complications. Current archiving practice requires a measure of active involvement in the preservation of electronic records, due primarily to the relatively fragile nature of the magnetic storage medium and to the need periodically to recopy records to meet changing industry standards, such as the density of magnetic tapes. This form of involvement can be expected to diminish significantly as a result of the adoption of optical disks as a storage medium, since they do not require the same attention to environmental conditions, they dramatically increase storage capacity; and they reduce costs as a consequence of the much greater data transfer rates of speed. The new media promise to spawn new demands for active involvement of archivists, however, especially in the need to copy current holdings on to newer media,[10] and to deal with the continued lack of agreement on technical standards. The question of information technology standards touches every archival function: appraisal, acquisition, processing, conservation and dissemination. Archivists may have to move beyond being knowledgeable about changing technology to becoming more directly involved in promoting efforts to ensure wide adoption of common standards.

Finally, we note that a recurring problem in the preservation of electronic records has been the software dependency of the records. Despite the recognition of the archiving problem created by software dependency and our earlier comment on the importance of depositing flat files, full software independence is not always possible. Indeed, the problem is exacerbated by some trends in information technology, such as the compound electronic document, constructed from several different sources. At the extreme, in cases such as the "virtual document," consisting of a set of pointers to other sources, this dependency dissolves into the absence of a source document itself.

Perhaps the most significant overall consequence of the recent developments in the creation of electronic records is the tendency for the originators and initial users of documents to become responsible not only for their creation, but also for their control and destruction. At the same time, it is unlikely that these users have the sensibilities,

information, training or interest required to ensure that documents of corporate or state value are saved for review and historical analysis. The trend gives rise to our reconsideration of the role of the electronic data archive.

New Approaches to Archives of Electronic Records

The topic of new approaches to conservation could be addressed under any of the archival functions relevant to electronic records, since the interrelationship of the functions is also fostered by electronic media, whether it is appraisal, arrangement and description or dissemination. One significant advantage of electronic documents is that all technical considerations revolve around the medium. A further implication of these developments may be a fundamentally different concept of archival functions in which the notion of custody of the records is abandoned and the archivist concentrates solely on appraisal and dissemination. In this case, the acquisition of records would no longer be a primary function of archives; custody of the records would remain with their institutional source, which would also bear responsibility for their preservation. In other words, the functions of archivists would become primarily intellectual: identifying electronic records of value, establishing standards for the long-term preservation of the records, monitoring the development of new systems, developing finding aids, and disseminating information about the records to researchers.[11]

The revised concept has some apparent dangers, of course, particularly, in the implications of abandoning physical custody of the records. Whether originating units can be persuaded to act as long-term conservation units is an open question, and one about which we are not especially sanguine. The increasing capacity and reduced costs of electronic records storage work in favour of preservation, but the historical durability of the electronic record is less impressive than hard copy, even in the age of the shredder. In view of the development of improved communication systems, however, and given the prospect of on-line transfer of data to researchers, the approach could provide for intellectual control without demanding the technical expertise for on-site physical control and conservation of the records. Moreover, it is not a trivial consideration that the change would save the expense of actual records acquisition and preservation.

This new conception of archiving procedures may prove to be feasible, at least, in cases of major government and private-sector corporate agencies as a result of emerging technology, but more importantly, it may become necessary simply as a result of the increasing irrelevance of the notion of original source documents. The transition from a focus on custody of records to intellectual control by archivists parallels the possibility raised earlier, that a national directory of historical databases derived from traditional sources may be a viable option, in the absence of institutional repositories.

Conclusion

This paper considers two related topics with respect to the conservation of historical data in electronic form: the question of documentation and access to individually constructed historical databases, and the issues raised for archivists by the rapid changes in electronic data processing technology. With respect to the first, we reviewed the options for adequate conservation and dissemination of historical databases that do not have major institutional support, and offer a number of specific recommendations regarding

documentation standards. We then reviewed major recent trends in information technology and the challenges which they present for archivists acquiring electronic records, ranging from new software development and communications technology to the apparently quite radical implications of electronically-created records for the traditional notion of a documents archive. It is proposed that such changes require an altered conception of the primary function of archives: moving away from the custody of records towards appraisal and intellectual control.

Notes

* This is a report drawn from the work of the Data Archive Group of the York Institute for Social Research and a paper by Gordon Darroch and Sue Gravel, presented to the Canadian Historical Association, Victoria, British Columbia, May 1990. It was written in collaboration with David Bates, Anne Oram and John Tibert of the Institute for Social Research, York University.

1 The question of variant and shifting terminology among archivists and social historians arises here. The term "machine-readable data" still has some, though increasingly limited, currency among historians using electronic records. The term is no longer current among archivists. Archivists in the National Archives of Canada, Government Archives Division are no longer distinguished by media, as they once were. In this report we use the terms "historical data files" and "historical databases" to refer to data in electronic form collected by historians in the context of specific research projects. In some cases, these files are actually complex, multi-file, relational databases, as in the case, for example, of the SOREP (Inter-University Centre for Population Research) project based at Chicoutimi, or the Research Programme in Historical Demography at Université de Montréal. Many other projects have generated more simply structured electronic files: see "Historical Databases: The Canadian Experience," *Histoire sociale/Social History* 21, (November, 1989), p. 42. We thank a reviewer for drawing these terminological distinctions to our attention.

2 There is a growing selection of well-documented database software available. The various editions of D-BASE are probably best known. A more complex, relational database management system is the *INGRES/INGRID* system adopted by SOREP. *ORACLE* is an unusual database management system that deals both with fixed field data and with unstructured textual data. For various applications and reviews of recent software developments, see the journal *History and Computing*.

3 See, for example, M.A. Strong, S.H. Preston, H.R. Lentzner, J. Seaman, and H. Williams, *User's Guide: Public Use Sample, 1910, United States Census of Population* (Philadelphia, 1989); S.N. Graham, *1900 Public Use Sample: User's Guidebook* (Seattle, 1989); S. Ruggles, "A Public Use Sample of the 1880 U.S. Census of Population," *Historical Methods* 23 (1990), pp. 104-115; David I. Kertzer, *Handbook in Casalecchio History Project* (Brunswick, 1985); David I. Kertzer and Dennis P. Hogan, "Appendix," *Family Political Economy and Demographic Change: The Transformation of Life in Casalecchio, Italy 1861-1921* (Madison, 1989); Jan Sundin, "Demographic Data Base, Umeå," from "The Conference on Methods for Using Population Registers in Historical Research," Umeå University, Sweden (August, 1984); Gérard Bouchard, *Annual Reports, SOREP* (Inter-University Centre for Population Research), Université du Québec at Chicoutimi.

4 It may be noted that at these centres the archival and documenting practices for historical data sets are identical to any other, for example, the files from a contemporary public opinion survey.

5 Initially, the documentation of some files may not be sufficient to allow ready access and use, but this fact could be entered in the directory and remedial efforts eventually undertaken.

6 The misrepresentation may have resulted either from a specific sampling design: for example, a small ethnic group might have been intentionally over-represented in order to permit detailed analysis, or from unintended differences between a sample and the population, for which adjustment is made after the fact.

7 To date, most local area networks do not allow one to monitor the creation of electronic records, as in the case of traditional record-keeping systems, although there are now efforts being made to

develop software for the purpose. Presumably, the inability to track source information and editing changes in decision-making documents is a liability for organizational functioning, as well as for historical analysis.

8 For an example of the application of Standard Query Language to accessing remote databases through a Wide-Area Network for the purpose of historical analysis, see David Gilbert and Humphrey Southall, "Data *glasnost*: a User-Friendly System for Access to Research Databases across Wide-Area Networks," *History and Computing* 3, no. 2 (1991), pp. 119-128.

9 For example, the "compound" or "smart" document and geographic information systems (GIS) are being introduced into a number of planning and administrative contexts. The former entail the integration of voice records, databases, documents and images, obviously overstepping the traditional media boundaries. In GIS a variety of administrative and planning data are geographically plotted and visually displayed.

10 For example, ICPSR at the University of Michigan expects to copy its entire magnetic tape holdings on to disk in the near future.

11 Some archives are already taking the initiative in monitoring the preservation of electronic records, while leaving custody to their creators.