The Relevance of Social Science Standards of Data Base Availability for Social History

GORDON DARROCH*

IN THIS NOTE I discuss the relevance to social history of the expectations and practices regarding public access to data bases in survey research. My research for this topic falls considerably short of survey research standards of adequacy. It consists of a number of years of experience as the director of York University's Institute for Social Research, which is largely a survey research and data archiving unit, and recent conversations with a small number of survey researchers and social historians in Canada and the United States.[1]

I focus on the question of expectations and practices regarding dissemination of data bases, rather than on the related question of documentation. The latter is an important topic, but it has been addressed elsewhere.[2] First, I discuss briefly the release of social survey data and then turn to some observations on the relevance of the former for the dissemination of historical data bases.[3]

The Release of Data from Contemporary Social Surveys

The release of data from social surveys has become relatively standardized in recent years, at least for major projects supported by public funding. There are now strong expectations and commitments on the part of social scientists that data files created from surveys will be clean, adequately documented, and available very shortly after their collection. As a matter of

* Gordon Darroch is a professor in the Department of Sociology at York University.

1 I have freely borrowed from interesting conversations with Gérard Bouchard, Myron Gutmann, and Steve Ruggles among historians and from David Northrup and Michael Ornstein of the York Institute for Social Research. None, of course, are responsible for my peculiar renderings of their ideas.

2 See, for example, Gordon Darroch and Sue Gavrel, with David Bates, Anne Oram, and John Tibert, "Preserving Historical Databases and Facing Technical Change: Common Issues for Social Historians and Archivists", *Archivaria*, 34 (Summer 1992), pp. 288–297. There are also several excellent papers regarding documentation questions in the most recent issue of *History and Computing*, vol. 4, no. 3 (1993).

3 I confine my comments to survey research in social science, since it is the major form of research that generates data bases. Other forms are conceivable, such as documentary research.

routine practice, the data are now released as soon as they are available to the principal investigators, often within a year or so of the end of data collection. One even detects a hint of professional competition among survey researchers in this regard.

A major qualification should be noted. There are still many variations among academic survey researchers and smaller survey research units, as well as among private survey organizations, which contract for academic studies. The data from some major Canadian surveys conducted in the 1970s, for example, were not released for over a decade. This is increasingly uncommon, however. It is now simply not imaginable that the data from a large academic survey would not be prepared for release in a matter of one or two years, given the general expectations of the survey community.[4]

I am not aware of a full analysis of the intellectual history of these changing academic standards, though this might be quite interesting. I do have an impression, however, of the main patterns of change over the last 40 years. Few formalized practices existed regarding data dissemination in social science surveys from World War II through the 1960s. Major archives were unknown. As the computerized storage of data revolutionized the conduct of surveys from the late 1960s, the expectation that major, funded research projects would be archived and documented for public release developed rapidly. The transition is probably best marked by the establishment of the University of Michigan's Inter-University Consortium for Political and Social Research (ICPSR). The ICPSR remains North America's major centre for storage and retrieval of survey research and historical data bases. The fact that in the late 1960s ICPSR's data files were entirely disseminated as IBM card files serves as a reminder of the technological transition we have witnessed in 25 years.

In the 1980s, the rapid release of survey data was further encouraged by the development of computer-assisted telephone interviewing, dubbed CATI. CATI technology makes possible the centralized collection of survey data by linking the telephone stations of interviewers with video display terminals that reproduce an image of an interview schedule. Data are entered directly to a data base and a number of checks and verifications can be performed automatically. Other features allow much more complicated

4 It may also be noted in this context that, although some private polling firms in Canada have made nonconfidential surveys available over the years, there is an increasing inclination to deposit polling data with academic archives. The Decima survey organization, for example, now uses Queen's University as an academic depository, and there is a formal association between the Goldfarb Company and York's Institute for Social Research. One survey researcher suggested that this may be partially in response to requests for public disclosure that have been pursued under the Freedom of Information and Privacy Act. It is also probably a response to the increased methodological sophistication of survey research, which draws the academic and private agencies together. Academic research conducted by private survey organizations is released in conformity with academic standards.

survey interviews to be conducted than are possible by traditional paper and pencil methods. The fact that useable data are created almost immediately following the start of interviewing also strongly encourages the expectation of rapid documentation and release of survey data. Preliminary analysis by principal investigators sometimes takes place in the course of the survey process.

Expectations for rapid release of data in social science have been further fostered by the policies of major public funding agencies, such as the National Science Foundation (NSF) and National Institute of Health (NIH) in the United States and of the Social Sciences and Humanities Research Council of Canada. SSHRC's 1992 guidelines regarding survey research specify:

> The data collected in a survey supported by the Council are public property and not the property of the principal investigator. They must eventually be made available to other scholars. The Council acknowledges the right of the scholar who has conceived and carried out the survey to enjoy the exclusive use of the data for a certain time. It expects, however, that this will be for a maximum two years after the end of the data collection phase. It encourages investigators to make data available earlier if possible.[5]

This is essentially a contractual agreement for the early release of data. In my view, however, the recent practices of survey researchers are not mainly attributable to such formal requirements, but reflect the changing normative climate surrounding the conduct of survey research. The formal requirements have institutionalized and reinforced the practice.

SSHRC's conditions regarding historical data bases are not identical to those applying to survey research, however. Historical studies fall under more general guidelines regarding electronic data. In this case, the Council only stipulates that the data become public property and be made available for use by others "within a reasonable period of time".[6] I suggest that the specific expectations regarding survey research reflect less a difference in the character of the data than a difference in method: survey research tends to be structured in terms of early, well-defined data-collection stages, while historical projects are more likely to have more or less continuous data collection and revision. These are only differences in tendencies in the conduct of research, but they are of some importance.

It warrants note that institutional requirements other than the directives of funding agencies often impinge directly on the conduct of survey research, but less on historical research. Academic survey research is normally

5 Social Sciences and Humanities Research Council of Canada, *SSHRC Grants: Guide to Applicants* (Ottawa: SSHRC, 1992), p. 91.

6 *Ibid.*, p. 92.

conducted with a clear knowledge of the regulations set by university senates, other academic governing bodies, and by ethics committees. These rules generally oblige researchers to place their results in the public domain within a relatively short period. For example, in the case of York University, the period is two years after completion of the project, with earlier release encouraged. The meaning of the phrase "the completion of a project" is unavoidably ambiguous, but in the case of survey research there is usually a distinct data collection phase, after which data are expected to be prepared for release. Given the routine scrutiny of the ethics of research projects conducted with live subjects, which includes survey research, I expect these institutional conditions impinge more directly on other social science researchers than on historians, though no doubt the distinction is not intended.

In sum, a number of intellectual, technological, and institutional changes have converged to increase the expectation that most social survey data will be quickly released to the wider academic community. Some examples may serve to illustrate the emerging practices. A major national survey of Canadian attitudes towards the Charter of Rights and Freedoms was undertaken in 1987. With the exception of the results of a few unusual methodological features, the entire file was available for secondary analysis within a *few months* of completion of the survey, long before the principal investigators had published their first analysis. The entire 1988 Canadian National Election survey was released within a year of the termination of data collection. More recently, a survey of attitudes and voting intentions in the October 1992 referendum was designed as an early phase of a national survey of the 1993 federal election. The referendum data were released in October 1993, the month of the election, obviously well before the completion of the data collection for the national study. Given these precedents and practices of survey research, what are the implications for historical data base projects?

The Dissemination of Historical Data Bases

Social historians appear to be no less committed than survey researchers to the principle that data bases should be readily available for secondary analysis, especially if data collection has been supported by public funding. At the same time, however, two significant differences tend to make the ideal of dissemination more difficult to fulfil for social historical research than for survey research. First, differences often exist in the institutional context in which the data are collected and documented. For the most part, survey research is conducted by contracting the data collection, as well as aspects of project design, to a permanent academic survey unit or to a private organization. Principal investigators work closely with a full-time professional staff in the design and execution of the survey, but the survey organization normally assumes responsibility for technical documentation and often for archive and dissemination functions as part of its contractual obligations.

The institutional contexts of social historical projects are more variable.

A few have created and sustained institutional frameworks in the course of building and maintaining data bases, for example, the Saguenay project, the demographic history projects of the Université de Montréal and the University of Utah, the Philadelphia Social History project, and the University of Minnesota Social History Research Laboratory. Even in these cases, however, continued funding for infrastructure support is often more problematic and uncertain than in the case of academic or private survey units, which normally conduct studies for a variety of researchers funded by both grants and contracts.[7] Certainly in English Canada infrastructure funding for social history projects has been scarce.

Second, as suggested above, what constitutes the end of the data-collection phase is less obvious in many social historical projects than in survey research. It follows that the dissemination and documentation of files have been more complicated. The difference can be illustrated by reference to some specific cases.

A number of impressive social history projects in the United States in recent years have sampled from the nominative manuscript records of historical censuses. There are now census samples for 1900, 1910, and recently for 1880 that have added to the public-use samples available from the U.S. Census Bureau for the decades from 1940 to 1980.[8] The University of Minnesota project is now in the process of complementing these with large, public-use samples from all the extant nineteenth- and twentieth-century censuses on microfilm. In addition there is Guest's National Panel Study, tracing birth-cohort samples from the 1880 to the 1900 census.[9] Each of these historical projects tends to match the one-time-only data collection design of most survey research. It is no coincidence that they also have explicitly adopted survey research conventions, including the early documentation and release of data files. For example, the 1880 census study of the Minnesota Social History Research Laboratory announced the release of a subsample of 50,000 cases prior to the availability of the entire data base. The principal investigators have also invited active interest in the project among other social historians and have provided published notice of the availability of the data base and its basic documentation.[10] One reason for such a public face is explicitly to encourage wide academic participation in the project, both to justify its funding and to enhance

7 I do not wish to suggest that academic survey research units are free of funding problems. In fact, they tend to experience frequent "boom and bust" cycles in funded project work.

8 S. Graham, *1900 Public Use Sample: User's Handbook* (Seattle: University of Washington Press, 1979); Steven Ruggles and Russell R. Menard, "A Public Use Sample of the 1880 U.S. Census of Population", *Historical Methods*, 23 (Summer 1990), pp. 104–115; Michael A. Strong, Samuel H. Preston, and Mark C. Hereward, "An Introduction to the Public Use Sample of the 1910 U.S. Census of Population", *Historical Methods*, 22 (Spring 1989), pp. 54–60.

9 Avery M. Guest, "Notes from the National Panel Study: Linkage and Migration in the Late Nineteenth Century", *Historical Methods*, 20 (Spring 1987), pp. 63–77.

10 Ruggles and Menard, "A Public Use Sample of the 1880 U.S. Census".

prospects of further funding for this and similar projects. The latter motive, of course, aims to benefit the larger community of social historians by fostering a climate of active public-agency funding. It is an example worthy of emulation.

The general importance of survey research in the United States is one of the conditions that has encouraged American social historians to make the early documentation and release of complete data bases an explicit objective and a rationale for their studies. It is possible that such conventions have not been as influential in Canadian social history because survey research has been of less academic significance here. Nevertheless, there are Canadian examples following the survey research model. One is the Guelph University project, which transcribed the 1871 Ontario census of manufacturing data to machine-readable form.[11] Another is the release of samples of the 1881 and 1891 manuscript censuses for Vancouver Island and Victoria, B.C., by the University of Victoria's Public History Group.[12] In both cases the dissemination of the data bases has been a main feature of the research. The cases also illustrate how the adoption of practices pioneered in survey research tends to assume specific, delimited phases of data collection in historical studies.

Such focused data collection projects will remain the principal form of social survey research and, perhaps, a major form of social historical research. I suggest, however, that they may become less common in historical work than even in the recent past. They stand in contrast to continuing survey and historical studies.

In survey research, continuing projects are normally designed as panel studies, which consist of a series of surveys tracing the experience of individuals over time. There are, for example, the General Social Surveys conducted by the Institute for Social Research at the University of Michigan and those conducted by Statistics Canada. In each case, documentation and dissemination of the data are accomplished in stages, following each panel survey. It hardly needs noting that these are large projects, with funding commitments over the long term and with the *central objective* of disseminating data to many researchers. A number of social and demographic history projects share with panel surveys this central feature of continuing data collection. Despite the apparent similarity, I argue that these historical projects face rather different and more difficult problems of data dissemination than studies with a single, early data collection phase, or panel surveys.

The path-breaking Hamilton project serves as one example. From the outset, Katz and his colleagues were well intentioned about making data

11 Canadian Industry in 1871 Project, *Bulletin No. 5*, University of Guelph, 1992.

12 See, for example, Peter Baskerville and Eric Sager, eds., *The 1881 Canadian Census: Vancouver Island* (Victoria: Public History Group, 1990), and *The 1891 Canadian Census: Victoria, B.C.* (Victoria: Public History Group, 1990).

available for secondary analysis.[13] For a variety of reasons, however, the ambition was difficult to fulfil. First were the pioneering character of the project, the lack of institutional context, and the lack of documentation conventions. Second, and as complicating, was the fact that new data were added over much of the life of the project, beginning with the several nineteenth-century censuses of Hamilton and subsequently including assessment and city directory data, newspaper references, and, in the end, comparisons with census data from Erie County, New York.

Other pioneering studies began as seemingly one-time-only projects, but new phases of data collection have developed, as in the case of Foust and Bateman's sample of northern, rural households from the 1860 U.S. censuses.[14] Recently, Atack, Bateman, and Gregson have begun the work of extending this sample by linking the data to the 1880 population and agricultural censuses. They expect to link back to the 1850 and 1870 manuscripts as well.[15] Ultimately, the authors will provide documentation for the new files, but it is not likely to be their first priority, nor will it be immediately available. It seems reasonable to think that, as the personal, computer-based record linkage programs of this project and others like it become available, they will foster the addition of new nominative data to other existing files, further complicating the notion of a single, complete, and readily accessible file in each case.

The major demographic projects, such as the Saguenay Project or those of the universities of Montréal and Utah, present quite different complications for the dissemination of data in continuing projects. Like those in social science panel surveys, these data bases are available, though not simply released to other researchers upon request. In each case there is some process by which an application is reviewed in terms of academic merit and possibly other concerns, such as confidentiality and privacy in the case of reference to living persons. My preliminary inquiries suggest the review processes are not standardized among the units and may require some patience on the part of prospective secondary users, though there is no doubt about ultimate access in some form.

Finally, if some parallels exist between panel survey research and continuing historical studies, there is other historical research quite unlike survey research. This work, conducted over a long period, involves data collection and revision as a central feature. One such project, undertaken by Peter Knights, began in 1972 and culminated in the publication of the book

13 See Michael B. Katz, *The People of Hamilton, Canada West: Family and Class in a Mid-Nineteenth-Century City* (Cambridge, Mass., and London, England: Harvard University Press, 1975); and Michael B. Katz, Michael J. Doucet, and Mark J. Stern, *The Social Organization of Early Industrial Capitalism* (Cambridge, Mass., and London, England: Harvard University Press, 1982).

14 F. Bateman and J. D. Foust, "A Sample of Rural Households Selected from the 1860 Manuscript Censuses", *Agricultural History*, 48 (1974), pp. 75–93.

15 Jeremy Atack, Fred Bateman, and Mary Eschelback Gregson, "Matchmaker, Matchmaker, Make Me a Match", *Historical Methods*, 25 (Spring 1992), pp. 53–65.

Yankee Destinies nearly 20 years later.[16] The book announces that the data base constructed in tracing two samples of men drawn from the manuscript censuses of Boston in 1860 and 1870 was deposited at ICPSR, presumably at the completion of the project. The data were not routinely available throughout the course of the project, however, since Knights continued to trace the sampled individuals through virtually every conceivable nominal record to their deaths (in nearly every case). In effect, the data base was really never complete until the analysis was concluded.

Although few, if any, other projects are likely to follow Knights's precedent with such devotion, others are similar in terms of the open-ended character of the data base. In studies I have undertaken, based on samples drawn from the 1861 and 1871 manuscript censuses of Ontario and the 1871 census of Canada, the data bases have been documented and made available to others,[17] but they also have been periodically revised, through tracing select groups into other records, as a consequence of reviewing the quality of some of the original data (such as that from the 1871 manufacturing census), by amending the weighting of cases in the sample, and by deleting problematic records. Documentation has routinely fallen behind data collection, revision, and analysis.

Another example is Myron Gutmann's Texas Demography Project, which combines the greatly enlarged storage capacities of personal computers with new software to trace individuals through a wide variety of nominal records.[18] The study design and technology allow detailed analysis to be pursued at the same time as continuing data collection and revision. Here again, documentation and release of data understandably trail behind the ongoing data collection and analysis, though earlier portions of the data base have been disseminated. Given rapid changes in computing and software, there is reason to think that this project represents something of the face of the future, in which data collection phases tend not to be neatly severed from analysis.

This brings me to two final issues. First, could access to the data bases of continuing projects be more readily facilitated by systematically releasing

16 Peter R. Knights, *Yankee Destinies: The Lives of Ordinary Nineteenth-Century Bostonians* (Chapel Hill and London: University of North Carolina Press, 1991).

17 The project was designed and initially conducted with my colleague Michael Ornstein. See, for example, Gordon Darroch and Michael Ornstein, "Ethnicity and Occupational Structure in Canada in 1871: The Vertical Mosaic in Historical Perspective", *Canadian Historical Review*, 61 (September 1980), pp. 305–333; and Gordon Darroch, "Class in Nineteenth-Century Central Ontario: A Reassessment of the Crisis and Demise of Small Producers During Early Industrialization, 1861–1871", in Gregory S. Kealey, ed., *Class, Gender, and Region: Essays in Canadian Historical Sociology* (St. John's: Committee on Canadian Labour History, 1988), pp. 49–72. For the national sample, consult the Data Archive Librarian, Anne Oram, York Institute for Social Research, York University, North York, Ontario.

18 On the project's record linkage, see J. E. Vetter, J. R. Gonzalez, and M. P. Gutmann, "Computer-Assisted Record Linkage Using a Relational Database System", paper presented to the Social Science History Association, Minneapolis, Minn., October 1990.

various versions of a file, from preliminary to increasingly complex ones? Second, what can we say about the concern that early release of a data base entails the risk that secondary users will usurp analysis that principal investigators envisaged in their design, even if only vaguely?

Releasing various versions of a data base is essentially the practice of social survey panel projects and of the major demographic projects, despite the variations among them in the processes of dissemination. By default, it is the practice adopted in such open-ended studies as Gutmann's Texas Demography Project. Clearly, in principle it is an excellent idea; in practice there are complications to be faced. Primary among these are the effort and funding required to provide multiple forms of documentation stretching over a number of years. Especially in the case of projects sustained by one or a very few researchers, there is the burden of revising documentation and of either managing dissemination or regularly updating the data base held by an archive. In some cases providing new releases of data files may be a relatively routine matter; in others it will be relatively time-consuming. Will funding agencies permit application for support of these continuing efforts? Will academic reviewers recognize such applications as sufficiently meritorious to support them?

With regard to the question of limited access to data while principal investigators undertake analysis, I offer several observations. On one hand, it is obviously a reasonable concern of any researcher who conceives and conducts a study, as SSHRC explicitly acknowledges with respect to survey research. On the other, it is worth noting that among survey researchers this consideration has not normally prevented the early documentation and release of their data bases. There are at least two reasons for this relative lack of concern. First, it is actually rare that a secondary user is interested in and prepared to undertake the kinds of analysis envisaged by the principal investigators. Second, it is usually only sensible to release a file of the original variables, not those constructed for the purposes of analysis, since the latter often entail a commitment to particular theories or forms of analysis and the construction of related derived variables, or require reworking a portion of the data base.

Finally, the question of proprietary rights to data raises a question of the nature of written agreements between secondary users and principal investigators or a disseminating archive. I am not aware of any essentially contractual agreements that limit specified uses of either survey or historical data bases, once a secondary user has been granted access to the data. However, in its recently proposed citation rules for machine-readable data, the Canadian Committee for History and Computing notes that the creators of data bases can assume to set conditions governing access to the data.[19] More

19 See José E. Igartua, "Citation Rules for Machine-Readable Data in Canadian Historical Journals", *Newsletter of the Canadian Committee on History and Computing, Canadian Historical Association*, II (Autumn 1993), pp. 9–11.

directly, the ESRC Data Archive of the University of Essex routinely uses "undertaking forms" to secure agreement on the form that references to an original data base will take in any publication.[20] Presumably, similar agreements on the types and limits of secondary analysis may be secured by principal investigators, while still making data bases widely available. Perhaps making such agreements routine would further enhance the dissemination and analysis of existing data bases, especially of those in progress over a relatively long term.

Conclusion

In sum, there seem to be two main orientations in social history with respect to the dissemination of data bases. The first encourages closely following the practices of survey research regarding the early documentation and release of data. In this case, like surveys, the historical projects tend to have well-defined, initial data-collection phases, after which analysis proceeds. The second orientation is among projects in which data base construction and revision is a continuing feature. In panel surveys, such data-collection phases are treated as separate surveys for the purposes of documentation and dissemination. Despite a common commitment to encouraging secondary analysis, many continuing social history projects are less readily structured in this way, and relatively few have the resources for routine revision of documentation. A changing climate of opinion among historians that values and credits the production, documentation, and dissemination of data bases, as well as greater willingness on the part of funding agencies to support these functions, would much encourage their systematic public release and use.

20 Sheila Anderson, "The Future of the Present: The ESRC Data Archive as a Resource Centre of the Future", *History and Computing*, vol. 4, no. 3 (1992), pp. 191–196.