A Study of Census Manuscript Data for Central Ontario, 1861-1871 Reflections on a Project and on Historical Archives

Gordon Darroch*

I — INTRODUCTION

I was surprised, but pleased to be invited to participate in this panel discussion of Canadian historical databases. The surprise resulted from the fact that I had not thought we were at a stage in the construction of machine-readable historical archives in Canada to warrant a collective reflection on the national experience. In an important sense, I believe we have just begun to understand the prospects and the limits of analysis of historical machine-readable data archives in this country. On reflection, however, I see that it is a good time to take preliminary stock and to assess the directions of development. I am pleased to contribute, on the one hand, because it provides me an opportunity to share some experiences in the design of a regional study that employs unique samples of census data and, on the other, because I have developed some specific notions about our responsibilities and difficulties in collecting and archiving nominal historical records.

II — A CAVEAT

Before I continue, it is important to comment on my qualifications to participate in a discussion of historical databases. I can offer no advice as a technical expert in database construction, maintenance or management. Statistical packages aside, I do no programming and I rely heavily on experienced programmers to manage and manipulate my data files. It is relevant to the purposes of this panel that my reliance on others has NOT prevented me from being very familiar with the details of the structure and management of the files. I expect that I am privileged in my access to fine and informed programmers and statistical consultants. I think, however, there is a general lesson here; lack of technical expertise need not be a barrier to constructing and analyzing large historical data files. On the contrary, I have come to believe from experience that reliance on a layer of professional and technical assistance is a considerable virtue in this enterprise. To be specific and slightly embarrassed in what may be a good cause, my colleague in this study, Michael Ornstein, and I lost a significant portion of our carefully constructed data, at one point, by merely being so distracted from data management as to allow expiry dates on computer files to pass unnoticed. Unlike conventional libraries, computer centres, as you know, do not routinely inform users of "due dates". There is a happy ending to this tale. Our near obsession with maintaining copies of earlier files and "hard" or paper copies of the data itself

* Department of Sociology and Institute of Social Research, York University.

Histoire sociale — Social History, Vol. XXI, n° 42 (novembre-November 1988) : 304-311

allowed us to reconstruct the files with modest cost, though not without a deep breath. Since this episode, we have given the management of the files over to an experienced programmer whose practices of file management, back-up and documentation are more professional and routinized.

III — THE STUDY

The study has been conducted in two phases. The first was focused on all four provinces of Canada in 1871 (Ontario, Quebec, New Brunswick and Nova Scotia). The second phase is focused on a large region of Central Ontario in the 1861-1871 decade (a wedge of counties stretching from the middle of Lake Erie to the lower shore of lake Huron on the west; and, on the east, from about Port Hope, on the north shore of Lake Ontario north, to the southern tip of Georgian Bay). The study is based on samples taken from the nominal data of the census manuscripts of those early years.

Both phases of the data collection have unique elements. The first phase created a representative national sample of households from the last century that allows detailed analysis of a variety of characteristics of individuals and their households. We have reported some results in historical journals (Darroch and Ornstein, 1980, 1984a, 1984b). The Ontario phase has three unique elements. First, it is based on record linkage of very large samples of individuals drawn from the census manuscripts of 1861 and 1871 in which the samples are clusters of surnames. These clusters allow probability samples to be linked. Second, we created records for these individuals that nearly exhausted the information from all schedules of the censuses of those years, including household information, farm tenancy and productivity, and the data of the manufacturing censuses. In some cases, this required additional linkage procedures to attach information from more than one schedule to the same purported individual. I will provide just a sketch of the methodology involved in a moment. Third, the samples are drawn for a very large contiguous region of Central Ontario, representing about half the provincial population in the 1861-1871 decade.

IV — CONCEPTUAL ORIGINS

A comment on the conceptual origins of the data collection procedures is necessary in order to understand the character of the data files. The study originated from a reading of two new types of historical work that had emerged in the 1960's and early 1970's. One was the breakthrough in demographic studies, represented by family reconstitution, using parish records in precensus times (*see* Fleury and Henry, 1956; Willigan and Lynch, 1982). At the time, this was largely a European development, although it was closely followed and much developed by Canadian studies (for Canadian work, *see* the other papers in this issue). The other work was the distinctively American tradition of close community studies of social mobility, stimulated by Stephan Thernstrom's first work, *Poverty and Progress* (1964).[1] Both literatures were distinguished by the systematic attention given to recovering aspects of the lives of ordinary people in the past. At the core, this concern with *social* experience was the feature that mattered most to me.

1. For the wide influence that this initial, if quite limited, work had on social and urban history in the United States, see *Social Science History*, Special Issue, Spring, 1986.

In each case, two aspects of the work caught my attention: one quite general and one very specific. First, these studies showed that a systematic and empirically grounded social history could be built up from historical sources for the great majority of people who left no intentional traces or records. Second and more specifically, these studies all faced, unsuccessfully, the serious problem presented by the facts of migration. The problem of migration was this: there was a great deal more of it everywhere, in every era, than historians or demographers had conventionally imagined.

We are now very familiar with the facts of migration or "transiency" in past time, although I also argue that we have largely become accustomed to them, rather than accounted for them or understood their implications. But the point here is that widespread community level migration seriously complicated, and perhaps actually jeopardized, the central objective of the new methodologies, since only the stable population for the selected local communities were "at risk" of being studied: the very large numbers of migrants simply escaped the analytic net.

In part, of course, the problem of migration stems from the arbitrary nature of the civil or administrative units most often adopted as convenient sites for study: a small town, a parish or two, a city, or possibly a county or department. Moreover, the difficulties presented by migration and the limits of civil units to tracing individuals through historical records were exaggerated in early studies. Still, recent historical studies of migration underscore the general difficulty, since they are based on rare historical sources, such as continuous population registers (for example, Kertzer and Hogan, 1985; Hochstadt, 1986), on the unique U.S. Soundex indexes of surnames (Stephenson, *et al.*, 1978), or on formidably tedious procedures of tracking individuals through innumerable discrete records (Knights, 1971).

V — NATIONAL AND SURNAME SAMPLES OF THE NINETEENTH CENTURY CENSUSES

Although migration was only one among several related concerns of our study centering on social class formation and the household economy after mid-Century, some solution to the problem was necessary in order not to vitiate any other results. The solution was to combine a methodological sledgehammer with a methodological scalpel. The sledgehammer was simply to expand the area under study sufficiently to capture the large component of total migration made up of local and circular moves, despite the heavy flow of outmigration to the U.S. in nineteenth-century Canada. The scalpel was a sample design.

Our sampling problem was very specific. Of course, one wants an efficient strategy that yields representative probability samples, but at the same time, the strategy had to allow us to conduct systematic record linkage between censuses. If one draws conventional samples from two or more historical listings, taking every nth person or otherwise randomly drawing cases, then, one virtually eliminates the possibility of systematically linking records for the same individuals from the different listings. The random element of the samples, which ensures the sample is representative, also ensures that there is only the very slightest chance that any one individual will appear in both samples. In sum, one needs both to sample and to ensure that the samples are effectively closed populations, so that the same surviving individuals will appear in each.

Our solution was to devise a form of letter or surname sampling. A sample of surnames, of course, preserves the possibility of record linkage (at least for those who do not change their last names, as women unfortunately tend to do on marriage). Our design can be briefly described, although the actual procedure is rather more tedious.

The first task was to demonstrate that a random sample of surnames was, in fact, an adequate representative sample of the population itself. For this purpose, a large random sample of households was drawn from the microfilmed copies of the nominal manuscript census of 1871. The information, for all individuals in the sampled households, was transcribed to paper and subsequently keypunched.

I wish to emphasize that we were compulsive in insisting on virtually complete transcription of the nominal records. A historical file such as this one is almost certain to be collected only once; the range of questions that might be addressed to the file can never be imagined in advance by principal investigators. Selecting data and precoding other information will always limit the potential uses of the files. Indeed, our own changed emphases and conceptual interests have more than once proved the point.

The sample provided a surrogate "national population" from which letter samples could be drawn and to which they could be compared for a variety of characteristics and relationships. As well, it was clear that we had an unusual opportunity to supplement our methodological concerns with substantive ones: a relatively large stratified, random sample of a nineteenth-century national population provides for unique and very rich sociohistorical analyses. As for the original methodological objective, the results were consistently encouraging: the design effects of letter samples, which are technically cluster samples, were modest and the letter samples adequately represented characteristics of the population from which they were drawn.

In this stage of the study, we adopted a refined version of the idea of letter sampling. The national sample was used to divide all surnames appearing in the census of 1871 into a set of about 100 mutually exclusive clusters defined by Soundex phonetic codes, using the first letter of the surname and a phonetic classification of the next portion of the name. From these clusters, a random sample of surname clusters or pockets was drawn, stratified by the size of the surname groups.

The data collection for Ontario differed from the national sample in that all the information from the several schedules of the censuses was recorded for every member of the households (in 1861, the personal schedule is supplemented by information about manufacturing and industries and by the separate agricultural schedule; in 1871, there were nine full schedules, including agricultural, industrial and real estate censuses). Despite the obvious limits of a single historical source — in this case, underenumeration, misreporting, limited descriptive information, the political and social prejudices of the census as a state agency and the simple semiliteracy of some enumerators —, these data still provide a remarkably valuable source for the analysis of individual lives and life cycles, of literacy, of property ownership and productivity, of labour, land and household economies (for example, Darroch and Ornstein, 1984; Darroch 1988).

The last of the major steps in this research design was the linking of the individual records, between 1861 and 1871, for the region in question. After reviewing well-known computerized procedures, we chose to develop a combination of computer and manual linkage that is particularly suited to historical census records (regarding record linkage methodology, *see* Wrigley, 1973; Winchester, 1985).

We judged that our chosen record linkage method was appropriate for this database, despite the relatively large data files, because they were still physically manageable as paper copy and could be carefully reviewed record by record by a few individuals over a number of months. In fact, the judgement was correct, although the linkage was no mean feat: there were over 34,000 individual records selected in the letter sampling for Central Ontario in 1861, and over 40,000 in 1871.

Using alphabetically sorted surname lists, the linkage proper combined a complex set of decision rules regarding records that would be allowed to refer to the same historical person (for example, nativity could not vary between censuses; but age, more or less than ten years; religion and phonetic spellings of names could, within specified limits) with the pattern recognition capabilities of research assistants.[2] The decision rules emerged out of a reading of the linkage literature and from trials undertaken by the principal investigators. In all, some 16,000 records were considered "true" links, although they were coded to include a subjective estimate of the level of certainty, which can be used as a variable in analysis. Our estimate for the entire region puts the rate of linkage at about 55 percent of those at risk in 1861, taking account of mortality and, for women, marriage and name change.

VI — A NOTE ON HISTORICAL ARCHIVES

Three issues regarding historical data files have arisen for me in the course of describing this project. First, I noted in particular the importance of full transcription of original manuscript sources of any kind in order not to eliminate future, unexpected analyses of the data. On several occasions in my relatively recent venture into social history, I have been generously offered access to files collected by others, only to recognize that early project-oriented coding and data collection decisions ruled out answers to my questions. One's choice in such circumstances is either to return to the original data and reinvent a portion, at least, of that particular archival wheel or simply to turn to other work. In my view, those of us who systematically create files from historical records have an obligation to play the parts both of archivist and of researcher. This is a relatively new joint responsibility and one that some historical researchers, but by no means all, have taken quite seriously. The obligation, I believe, is not restricted to those of us who have had the privilege of public funding; it is a more general obligation to the community of social historians present and future.

Second, fulfilling the role of researcher and archivist requires planning and practices that frequently go beyond those of traditional social science or historical research, especially in the provision of access to machine-readable files for others to undertake their own "secondary" analysis. Others on this panel have experience with and views on the technical issues, in particular on the question of adequate documentation; I only wish to make a more general comment.

In recent Canadian work, it seems to me, there are two main types of historical files. The first have a relatively simple structure, such as those I describe here. I would not

2. We found that some individuals were much more adept than others at retaining and recognizing fairly complicated combinations of name spellings, personal and household characteristics, while searching the "pockets" of similar names for matching records. A number of systematic, independent searches were made on portions of the data to compare results of different individual decisions.

normally call them databases, although I know the term is a very general one. They may, in fact, be large and contain very many variables, but they are simple in that they are "rectangular" files with a constant set of variables describing a well-defined and limited set of entities — persons or households, for example. The other files are true databases. As units of analysis and file organization, they tend to have several or many kinds of entities; for example, persons, households, areas and relations among these. They are constructed as relational systems, linking, perhaps, a number of separate rectangular files. The management of these databases is very different from the management of simple ones and much more costly in time and energy. I am venturing to the limits of my knowledge here, but I make the distinction to point out that these two types of historical data raise quite *different* archival issues.

In the first case, the files are unusually constructed and managed by a single investigator or small number of researchers. Community studies tend to create simple files. The archival issue in this case, I believe, is about the abilities and willingness of individual principal investigators to make the files clean and accessible and to provide sufficient, if not full, documentation for potential secondary users. My experience with both historical and social science data in Canada is that most researchers are relatively poor and uninterested archivists. Our data has been made available in several forms to a few other researchers, but our documentation is minimal and I must confess, daunting to them. The important Hamilton project, for example, suffers from similar limitations (Katz, 1975; Katz, Doucet and Stern, 1982).

The archival problem — the problem of open access — is different, I presume, for the complex, relational databases, such as the Saguenay project. I am guessing, but in these cases, the projects must employ (or principal investigators must be) a specialist in designing relational databases. The main problem for the secondary users, then, is normally not limited documentation, but how to understand the documentation that will be routinely (or should be) generated by the specialist. Specialized professional practices, especially computer-related ones, have a tendency to create their own argot, which is never especially inviting to the uninitiated. One question strikes me as an appropriate guide for adequate database documentation. Can one tell fairly quickly, *by reading prose,* if the data will answer one's historical questions? I argue that we have an obligation to communicate widely and in ordinary, nontechnical language to ensure a heritage of accessible nominal historical data.

VII — REGIONAL PERSPECTIVE

Third, and finally, I wish to comment on what I now see as the most valuable, substantive contribution of a regional study based on single source nominal data files, such as the ones I have participated in creating. There is some sense, I believe, that the initial pathbreaking phase of social, demographic and labour history has passed. In Canada, there are still relatively few, if very important, close studies of local communities systematically employing nominal data in conjunction with other more traditional sources (for Ontario, Katz, 1975; Gagan, 1981; Akenson, 1984; Gaffield, 1987; for Quebec, *see* references in Bouchard and in Charbonneau, in this issue).[3] Their still small numbers, I believe, makes

3. Despite recent lively and valuable contributions to Canadian labour history, neither this work nor social history, more generally, has much developed the use of systematic databases for the analysis of class and of labour in Canada. But, *see* Igartua, 1987.

them bear an inordinate burden of interpretation. Moreover, despite the depth, the nuance and insight with which local studies bring to their subject, they are not able, normally, to surmount its social and geographic limits. There are too many pieces to the puzzle cut this way alone. One consequence is that we lack a broader perspective on the institutional landscape and its relation to individual lives. Certainly, this is true for nineteenth-century Ontario on which my project is centred. How did the individual and family experiences and structural patterns of Hamilton, Peel, Leeds and Lansdowne and the Ottawa Valley reflect and contradict the larger social and economic landscape?

There is a current need for greater attention to synthetic and integrating studies and, to some extent I think, for a revitalized intellectual agenda in Canadian social and labour history. In this context, I suggest there is a unique contribution to be made by specifically regional studies, one form of which is founded on regional databases. Such studies are one, empirically grounded means of broadening perspectives on the relations between individual experience, patterns of institutional variation and wider structural change.

REFERENCES

Akenson, Donald. *The Irish In Ontario: A Study In Rural History.* Kingston and Montreal, McGill-Queen's University Press, 1984.

Darroch, Gordon and Michael D. Ornstein. "Ethnicity and Occupational Structure in Canada in 1871: The Vertical Mosaic in Historical Perspective", *The Canadian Historical Review*, 61:3, September 1980, pp. 305-333.

Darroch, Gordon and Michael D. Ornstein. "Family and Household in Nineteenth-Century Canada: Regional Patterns and Regional Economies", *Journal of Family History*, 9:2, Summer 1984a, pp. 158-177.

Darroch, Gordon and Michael D. Ornstein. "Family Coresidence in Canada in 1871: Family Life-Cycles, Occupations and Networks of Mutual Aid". Canadian Historical Association, *Historical Papers*, Vancouver, 1983, 1984b, pp. 30-55.

Darroch, Gordon. "Class in Nineteenth-Century, Central Ontario: A reassessment of the crisis and demise of small producers during early industrialization, 1861-1871", *Canadian Journal of Sociology*, 13:1-2, September 1988. Gregory S. Kealey, ed., *Class, Gender, and Region: Essays in Canadian Historical Sociology.* St. John's, Committee on Canadian Labour History, 1988.

Fleury, Michel and Louis Henry. *Des registres paroissiaux à l'histoire de la population : Manuel de dépouillement et d'exploitation de l'état civil ancien.* Paris, Éditions de l'I.N.E.D., 1956.

Gaffield, Chad. *Language, Schooling, and Cultural Conflict: The Origins of the French-Language Controversy in Ontario.* Kingston and Montreal, McGill-Queen's University Press, 1987.

Gagan, David. *Hopeful Travellers: Families, Land, and Social Change in Mid-Victorian Peel County, Canada West.* Toronto, University of Toronto Press, 1981.

Hochstadt, Steve. "Urban Migration in Imperial Germany: Toward a Quantitative Model". Paper presented at the Canadian Historical Association meeting, Winnipeg, June 1986.

Katz, Michael B. *The People of Hamilton, Canada West: Family and Class in a Mid-Nineteenth-Century City.* Cambridge, Mass., Harvard University Press, 1975.

Katz, Michael B., Doucet, Michael J. and Mark J. Stern. *The Social Organization of Early Industrial Capitalism*. Cambridge, Mass., Harvard University Press, 1982.

Kertzer, David and D. Hogan. "On the Move: Migration in an Italian Community, 1865-1921", *Social Science History*, 9:1, Winter 1965, pp. 1-24.

Knights, Peter. *The Plain People of Boston 1830-1860: A Study In City Growth*. New York, Oxford, 1971.

Stephenson, Charles *et al. Social Predictors of American Mobility: A Census Capture-Recapture Study of New York and Wisconsin, 1875-1905*. Chicago, Newberry Library, 1978.

Thernstrom, Stephan. *Poverty and Progress: Social Mobility in a Nineteenth- Century City.* New York, Atheneum, 1964.

Winchester, Ian. *Record Linking in the Microcomputer Era: A Survey.* Umea-Haparanda, Demographic Database Newsletters, 3, iii, 1985.

Willigan, J.D. and K.A. Lynch. *Sources and Methods of Historical Demography.* New York, Academic Press, 1982.

Wrigley, E.A. *Identifying People In the Past*. London, Edward Arnold, 1973.