Lisa Y. Dillon

Canadian Historical Association Annual Meeting, Ottawa, May 30-June 1, 1998[1]

Draft #2

Guidelines for the Creation of Historical Microdata

Historians interested in the historical experiences of ordinary Canadians face a considerable challenge. Letters and diaries, company and association reports and minutes, government records, and contemporary newspapers, magazines and literature shed light on prescribed ideals, social discourse, and, to a certain extent, actual behaviour. However, since the 1960s, historians have drawn upon routinely-generated sources such as parish registers and census enumerations to gain a long-term and systematic perspective on individual and family behaviour.[2] The major challenge faced by these historians has been to transform such extensive and detailed sources into the more manageable form of machine-readable data. The creation of historical microdata allows researchers to take advantage of computer processing and statistical analysis to study the interrelationships among individual and family characteristics. The purpose of this paper is not to assert the validity and usefulness of such projects, which has been debated elsewhere, nor to describe the broad research agendas within which they must be situated. Rather, this paper offers practical guidelines for conducting historical microdata projects, focusing on census projects in particular.

Canada's historians have been part of numerous historical microdata projects. Examples include the national sample of the 1901 Canadian census created by the Canadian Families Project, University of Victoria; the national database of the 1871 Canadian census created by Gordon Darroch and Michael Ornstein at York University; the coded production records of all industrial establishments in the 1871 industrial census by Kris Inwood and Gerald and Elizabeth Bloomfield at the University of Guelph; and the vital events database for residents of present-day Quebec from 1608-1850 by l'Institut Interuniversitaire de Recherches sur les Populations.[3]

Outside Canada, the Minnesota Historical Census Projects at the University of Minnesota, the Nineteenth-Century Censuses Collection Project at the University of Essex Department of History and History Data Service and the Geography Department at Queen Mary and Westfield College, University of London, and the 1801 Census of Norway project at the Department of History, University of Bergen are three further examples of ongoing projects to create historical data.[4] Initiating and running substantial projects like these requires much planning and organization. Historians involved in historical microdata projects know and communicate with each other. Indeed, many articles and research notes have been published in journals such as Histoire Sociale/Social History, Historical Methods, History and Computing, and Computing and the Humanities which examine the use of historical microdata for historical research. Such articles have explored the range of theoretical questions which can be addressed through quantitative analysis, critiqued the collection procedures and assumptions characteristic of various routinely-generated sources, discussed the challenge of integrating the results of quantitative research into the broader narrative of Canadian history, and suggested continuities and discontinuities between qualitative and quantitative research.[5] Other articles have provided detailed descriptions of the databases themselves, reviewing sampling and linking strategies, explaining the coding of variables and articulating the advantages posed by new software and hardware, while still others have explored how to use sophisticated statistical techniques in the course of quantitative research.[6] However, historians' discussions of how to plan, fund and run historical microdata projects are usually informal. To articulate formal guidelines for managing historical microdata projects, this paper draws upon the experiences of the Canadian Families Project (CFP) and the Minnesota Historical Census Projects.

Since 1996, the CFP at the University of Victoria has been conducting one of the most large-scale and comprehensive research projects undertaken by historians in Canada in recent years. The CFP is a Major Collaborative Research Initiatives project (MCRI) funded in part by the Social Sciences and Humanities Research Council of Canada (SSHRC) to create a machine-readable database of the 1901 Canadian census. The CFP team is composed of scholars drawn from across Canada, who will use the 1901 census database to study Canadian families. The Minnesota Historical Census Projects encompass a variety of projects conducted by the Social History Research Laboratory (SHRL) at the University of Minnesota.[7] Beginning in 1988, the Minnesota Historical Census Projects created a 1-in-100 public use microdata sample (PUMS) of the 1880 U.S. census. Since then, the Minnesota project team has created national PUMS of the 1850 and 1920 censuses, and is currently working on PUMS of the 1860, 1870 and 1900 U.S. census. The Minnesota Historical Census Projects has also modified existing PUMS of twentieth-century U.S. censuses so that the entire series from 1850 to 1990, with the exception of 1890, can be analyzed together.[8] The Canadian Families Project and the Minnesota Historical Census Projects have faced many similar challenges in the course of their work. Examining these issues yields important advice on how to apply for funds, draw up a budget, plan project infrastructure, account for computing needs, identify appropriate computing consultants and programmers, plan data entry procedures, hire data entry assistants, and create dictionaries and codebooks to describe the microdata.

Application Process and Budgeting

The first stage in the creation of historical microdata must be the existence and articulation of broad research goals. As David Gagan argued in 1988, creating a large database first and identifying research opportunities afterward is a haphazard approach. Instead, he stated, quantitative social historians must be "armed with an agenda derived from the compulsion either to refashion the outlines of Canadian history or to integrate the common and the unique aspects of Canadian social development into the wider canvas of the social history of western civilization."[9] The Canadian Families Project was inspired by broad questions about the Canadian family such as whether there has ever been a "traditional family," how we acquire our various ideals and definitions of family, and what living standards were like prior to the emergence of the welfare state.[10] Canadian researchers interested in exploring similarly broad research questions through analysis of a routinely-generated source should then consider with whom they would like to collaborate. As explained below, collaborative endeavours probably offer the best chance of success in managing a large-scale historical data project.[11] Once a group of scholars have agreed to collaborate in a historical microdata project, they must decide who will serve as project manager and which institution shall host the stage of data creation. They must then make preliminary inquiries to determine if this host institution is interested in and willing to support their research project. These steps are a necessary preparation to the crucial fifth step: obtaining funds.

For major research projects, the most common funding agency Canadian historians turn to outside their own universities is SSHRC. In the United States, most researchers interested in creating historical data apply to either the National Science Foundation (NSF) or the National Institute of Child Health and Human Development (NIH). For good reasons, SSHRC will only fund the creation of historical microdata if these data are conceived as a tool to enable investigators to address clear research goals. Assuming that most quantitative historians start with specific research goals in the first place, the challenge faced by the collaborative team is therefore to articulate how data creation will facilitate their research. The CFP did so in its MCRI application by first defining a variety of research goals centering on the family and imbedding within those goals a phase of data creation. Thus, the CFP is not simply a database project. While the first part of the CFP project consisted of data entry, cleaning, checking and coding and the creation of a codebook describing the data, the second phase of the CFP consists of research by the various CFP team members using different aspects of this database. The CFP tailored its application to emphasize that the project output would appeal to the scholarly community, as well as the wider public. To disseminate research results to the academic community, the CFP MCRI application stated project members would write and present working papers, conference papers and scholarly publications. It proposed to reach out to the broader public through a historical version of the Vanier Institute's Profiling Canada's Families (1994), as well as a CD-ROM on Canadian families.

In contrast to the Canadian experience, the U.S. NSF and NIH funding agencies prefer grant applications which focus on creating bodies of data. These granting bodies are very specific about not wanting research plans attached to database projects, as they want the resulting databases to be flexible enough to benefit a range of researchers. Nevertheless, the principal investigator of the Minnesota Historical Census Projects, Steven Ruggles, has been careful to include in his applications long-term objectives which articulate research areas which will benefit from the creation of U.S. census data in particular years. The difference between the CFP and Minnesota Census Projects' applications is that the Minnesota Projects budget for data entry and processing only, and do not request funds for subsequent research.

The fact that the CFP consists of an inter-university team of historians, two sociologists and one geographer rather than one or two scholars based at one university was another key element in the success of their application. Other Canadian researchers who wish to fund a project at a very large scale should think in collaborative terms. A single person would not have the skills and time to be responsible for such a project on his or her own. Fortunately, such databases are certain to be of interest to historians at other universities as well as researchers in other disciplines. In addition, today's internet technologies such as electronic mail, the world wide web and file transfer protocol facilitate the collaboration of nationally- and internationally-dispersed scholars. Work load aside, it is very unlikely that a researcher could win sufficient funds to create a large-scale set of historical data if he or she applies to SSHRC as a single person. SSHRC requires assurance that microdata created in the course of a grant benefit a multitude of users; writing an application as part of a team of researchers is a key way to respond to this requirement. The CFP put a great deal of thought into the collaborative nature of its project. Its proposal presents separate and very specific discussions of each team members' individual project, including descriptions of which additional primary sources would be used to shed light on issues which a census database cannot address. To highlight its integrated research approach, the CFP also made explicit the links among these individual projects. One SSHRC official later commented to the CFP that this project seemed truly to bring together a group of scholars; SSHRC detected in the CFP application an "esprit de corps" which was missing from many others.

The CFP wrote an application which adhered as closely as possible to the SSHRC guidelines. "When SSRHC states that student training must be a part of your project, they mean it," states

Eric Sager, principal investigator of the CFP. Undergraduate and graduate students, as well as post-doctoral fellows, can be involved in historical microdata projects from beginning to end, as data entry operators, as checkers, as variable coders and as analysts. Microdata lend themselves to varying levels of student analysis, from fourth-year honours papers to doctoral dissertations. Grant applicants should also seek advice from scholars who have already won such grants, rather than write their applications in isolation. The letter of intent required in the first stage of applying for a SSHRC MCRI grant is a good case in point. Within only eight pages, project collaborators must detail the objectives of the research project, the nature and extent of the collaboration to be undertaken, the proposed budget, the preliminary research plan, and a preliminary plan for communicating research findings, among other items. It is only after this letter of intent has been accepted that project collaborators can proceed and submit a formal application for an MCRI. Given the importance of the letter of intent, investigators of new projects should examine the letters of intent from past applications before writing their own.

The CFP project made a serious effort to write an accurate budget from the start, and had to examine carefully a whole series of issues when preparing it. They turned to the University of Victoria's Directorate of Research, Computing Services and local computer stores when estimating costs. When the CFP prepared its budget for SSHRC, it listed all project costs, then listed other available funding, including institutional contributions, and then specified the total requested from SSHRC. A summary of items included in the CFP and Minnesota Census Project budgets is below. Each of these two projects included these items somewhere in their budgets:

Personnel:

- Principal Investigators/Research time stipends[12]

- Programmer
- Data entry Operators
- Research Assistants[13]
- Consultant Costs (Professional/technical services)

. Equipment (workstations, microfilm reader, microcomputer)

. Supplies (tapes for data back-up, diskettes, laser printer supplies, miscellaneous office supplies)

. Travel (to release data at conference, publicize the sample & obtain feedback from users)

. Infrastructure costs of networking activities

. Other Expenses (purchasing microfilm of census enumeration forms[14], equipment maintenance, photocopying, long-distance telephone postage and page charges)

On the one hand, it is important not to pad the budget unnecessarily, since assessors are required to comment on credibility of the proposed budget. On the other hand, it can be an equal mistake to give yourself short shrift. In hindsight, the CFP principal investigators now believe that they should have requested more funds for computer consulting costs, conference travel, doctoral research assistants and post-doctoral fellowships.

Grant applicants can advance their cause considerably by referring to past successes with historical microdata. Here, a team approach poses an advantage by allowing grant writers to describe the experience of several historians with machine-readable history data. Some of the Minnesota Projects' current success with grant applications is a result of their admirable past record. In his application to create databases of the 1860 and 1870 U.S. census, Ruggles noted that "Experience…on the three previous census projects completed by the investigators will result in significant cost efficiencies for the 1860 and 1870 PUMS. If this project is undertaken now, it can be completed at a lower cost than any previous national census microdata file."[15] Ruggles also notes in his applications that he has performed pilot studies of the proposed project, funded through other means, to test its viability. In their own application, the CFP described the past experiences of its co-investigators with large-scale databases through, for example, the Atlantic Canada Shipping Project, the Hamilton Social History Project, the Canadian Historical Mobility Project, the Rural and Small Towns Research and Studies Programme, and the Saguenay Project. The CFP referred to these past projects not only to demonstrate the CFP members' experience with historical microdata, but also to discuss the broader analytic opportunities posed by creating a machine-readable version of the 1901 Canadian census. The CFP stressed that studying results from the 1901 census in conjuction with results from these other samples would allow its team members to examine change over time.

Infrastructure

Commencing and running a historical microdata project requires more than just funding; it also requires space and infrastructure. Unfortunately, the first big problem in universities today is space: requests for space seem to alarm administrators more than money questions. The cost of computers is nothing compared to lighting, floor space, windows and ventilation. For this reason, it is crucial for principal investigators to obtain a promise of space from their university administration before they submit a grant application. The CFP's acquisition of L-Hut, a war-time barracks converted to office space, was considered a campus coup. In their SSHRC application, the CFP described the University of Victoria's contribution to the CFP budget as host institution. They specified that this contribution included hydro, heat and personnel costs, as well as funds for computing, programming, statistical consulting, research release time and cartographic work. The CFP application also stated the total dollar contribution by the other five universities associated with this project, "variously allocated to computing, computer programming, travel and statistical consulting." In the early years of the Minnesota Census Projects, Steven Ruggles had difficulties obtaining sufficient space for his large staff of data entry operators and research assistants. He eventually expanded his space from one room, sized 14 1/2 by 19 feet, to three offices, plus extensive cupboard space for microfilm reels. Ruggles did so by winning the sympathies of the department chair and by promising relocated history professors that he and his research assistants would move their office belongings themselves. Including Ruggles' own office, the Minnesota Census Projects now occupies one-quarter of one floor in the University of Minnesota's Social Sciences Tower. Drs. Sager, Baskerville and Ruggles successfully argued for further space on the basis of their projects' value to their departments of history and to the Universities of Victoria and Minnesota. Unlike the U.S. funding agencies, SSHRC will not provide for spatial infrastructure. SSHRC requires that research applications be accompanied by a written assurance from the host institution that it will allocate space and infrastructure to this project.

Researchers must also determine in advance what their computing needs will be and who will provide consulting services. The computing needs that arise in the course of a historical microdata project are numerous, even setting aside data entry. Who will select and purchase your computers? Who will install the hardware and who will maintain it? Who will connect your computers to network ports so that they are networked to your university's mainframe and the internet? Some of these initial tasks can be performed by a university computing services' hardware specialist. However, one cannot assume that local university computer services will be able to respond to all the needs that crop up in the course of a historical microdata project. Historians' experiences in this matter vary widely. For instance, the Memorial University of Newfoundland was able to lend more routine support to the Atlantic Canada Shipping Project because this project was very large and because it featured a couple of very prestigious researchers on its team. The Minnesota Historical Census Projects relies on the University of Minnesota Social Science Research Facilities Center (SSRFC) to provide a fileserver, long-term data back-up, an ethernet system and other computer network support. It also turns to the Department of History's computer laboratory co-ordinator to help it select and set up computers and install software on those computers. The microfilm readers and simple repairs to them are done by the data entry operators on the Minnesota projects, although they prefer that a professional be hired to clean them at regular intervals. In contrast to the Minnesota Historical Census Projects, the CFP found that the University of Victoria Computing Services staff deal only with permanent institutions, and cannot be deployed on an ad hoc basis for special projects. While the CFP garnered a university technician to help them set up their computers at the beginning of the project, they have had to find alternate computer support in the course of the project.

If your university cannot provide routine hardware and software support, the researcher must budget and use some of their SSHRC funds to hire private computer consultants. This, too, is a tricky process, partly because it can be expensive but more because it is difficult to find a consultant who is sensitive to the needs of historians. Even the Minnesota Historical Census Projects, which could turn to a variety of university resources for hardware and software assistance, had to find and cultivate the skills of a person who could both understand and program historical data. In this instance, Steve Ruggles found out that a Ph.D. student in the department of history, Todd Gardner, had an undergraduate degree in Physics and a background in data processing. Since 1988, Gardner has written all the data entry programs for the Minnesota Census Projects; he wrote these programs in C language. With the 1860 and 1870 census project, the Minnesota Projects are switching to the database program, ACCESS, to process its census data, but it is still using data entry programs written by Gardner. Since Gardner was at the same time completing a Ph.D. thesis based on the Minnesota historical census data, he had a special interest in his computing work. The Minnesota Projects' switch to ACCESS for post-entry processing (and, eventually, data entry itself) is notable: with the improved performance of computers, more flexible and powerful software has come on the market. The availability of ever-more sophisticated hardware and software becomes an issue for ongoing historical microdata projects. Ruggles and the Minnesota Projects graduate assistants have had to confront questions such as "Should we buy new computers now or wait? Should we develop new software now or stick with what we've got?"

The Canadian Families Project has had a somewhat parallel experience. In the early stages of this project, co-investigators Eric Sager and Peter Baskerville hired Todd Gardner to adapt the Minnesota data entry software to the needs of the 1901 Canadian census. Gardner visited Victoria for one week to install this software on the CFP computers and demonstrate its use. Hiring Gardner bore some advantages, in that it was relatively inexpensive to have Gardner transfer technology he had developed for another project to this one. However, having a programmer write the data entry software in a programming language and then leave meant the CFP had less flexibility in changing the programs afterward. For a time, Dr. Gardner communicated with the CFP via e-mail and adjusted their programs as needed. As the CFP's needs for software and programming support grew, however, they hired two additional staff members, Doug Thompson and Marc Trottier. Thompson managed the data entry operators and performed most software management. When the CFP decided to enter Schedule 2 alongside Schedule 1 of the 1901 census, Marc Trottier, an engineer, was able to use more sophisticated software to devise a secondary data entry program. Using a combination of PARADOX and SPSS software, Trottier wrote additional programs to check, clean and verify the data. Now that the data entry phase of the CFP is complete, Trottier is chiefly occupied with writing SPSS programs to construct new variables on the basis of old ones. Although Trottier had been trained in PARADOX, he picked up and wrote programs in SPSS so they could be understood by the project investigators.

Todd Gardner, Doug Thompson and Marc Trottier are all unique assistants who combine a talent for computers and programming with a sensitivity to historical issues. Both Gardner's presence on the Minnesota Historical Census Projects, and Thompson and Trottier's association with the CFP occurred somewhat through happenstance. Unlike the other graduate assistants working on the Minnesota Historical Census Projects, Todd Gardner happened to have a background in computer programming. However, project investigators cannot assume their Ph.D. students have technical knowledge, and the CFP did not have any graduate students to fulfill the role Dr. Gardner played for the Minnesota Projects. Instead, the CFP came upon Doug Thompson and Marc Trottier when funding to support their work elsewhere in the University of Victoria's Department of History came to an end.

In summary, it is very rare to find a history graduate student with extensive technical knowledge, and universities themselves cannot always provide this support on a day-to-day basis. Yet, history microdata projects need consulting support from someone who can both program and understand what a historian wants to do. In fact, as Gordon Darroch and Michael Ornstein found in their management of census data for 1861-1871 Ontario, "reliance on a layer of professional and technical assistance is a considerable virtue in this enterprise."[16] An experienced programmer can bring to a historical microdata project more professional and routinized practices of file management, back-up and documentation.[17] Therefore, project investigators must budget at least 6 months before they begin entering data to find a satisfactory consultant and prepare for data entry. When interviewing potential consultants, investigators should start by having them examine the routinely-generated historical source that will form the basis of the microdata project. Members of the Atlantic Canada Shipping Project had met many programmers before they interviewed Wilf Bussey, who worked for Memorial University. Bussey wanted to know what their research questions were and requested a list of the variables they would be examining in their data. When Bussey asked if he could take away a box of crew agreements for a week before submitting a consulting proposal, the project members knew they had met the right consultant. It is crucial that the hired consultant understand the challenges the project investigators are facing by taking in the structure and meaning of the routinely-generated source, for example, seeing that some of the data is alphabetic and some numeric, that household-level characteristics must be linked to individual records, and that hierarchical files must be rectangularized. As Dr. Sager notes, "If you find a computer person who doesn't want to sit and read your source for a few hours, you haven't got the right person."

Data Entry

Procuring the equipment necessary for data entry requires much thought and preparation. Grant applicants should be careful to budget for paper, photocopying and telephone bills; these items are covered by SSHRC. Since SSHRC will not provide funds for telephone installation and furniture, researchers must ensure that their host institutions will provide telephones, chairs, tables and microfilm readers instead. Dr. Sager notes that it is important to pay attention to the ergonomics of data entry; most universities will have a staff member who knows something about the working conditions necessary to perform the arduous task of data entry. It is crucial to have chairs with adjustable up and down mechanisms and to have footrests and tables of the correct height. These chairs should also have good and adjustable back support; some data entry operators may also be helped by arms on the chairs or other ergonomic arm supports. Project investigators will probably find they must make changes in project set-up as data entry gets under way, and should decide in advance who will deal with these matters. In Minnesota, Ruggles has been able to secure sufficient funding to delegate this task to a graduate student research assistant, who in turn orders equipment upon the prompting of the data entry operators. The CFP, however, will not take on graduate students until its research phase; as a result, the two co-investigators supervising the data entry phase, Drs. Sager and Peter Baskerville, have had to deal with equipment problems themselves.

The Minnesota Census Projects' structural arrangements for data entry have changed considerably over the years. At first, four data entry operators and one research assistant worked together in one room. These close quarters greatly hindered the operators' abilities to establish a comfortable work routine. Their need for greater space provided a major impetus for the

Minnesota Projects' subsequent spatial expansion. The CFP encountered considerable difficulties with microfilm readers. Most historical microdata projects will not need microfilm readers which also print. However, since libraries now almost universally order reader-printers, microfilm machines which only read are very hard to find. On one brand of microfilm reader, DUKANE, the bolts for the turning mechanism fell apart and could not be replaced. When the CFP ordered the EYECOM microfilm readers, which feature both fiche and film-reading mechanisms, they were delivered very late. At $1,200 each, these machines are also very expensive, with no market for resale.

Research planners must also consider who will perform data entry on their projects, and whether their funding source allows them to hire professionals or students. The Minnesota Census Projects employs professional data entry operators, all of whom brought past data entry experience to their job and most of whom also have a liberal arts background. SSHRC dictates that students be hired for all research assistant positions. Drs. Sager and Baskerville advertised the data entry positions with the CFP and were able to find students in the humanities and social sciences with previous data entry experience. History professors are rarely in the position to hire data entry operators. However, most universities will have staff members who can advise researchers on how to interview students for this sort of staff position. The same interview pattern must be repeated for each research assistant, and it is important to school yourself against false first impressions. At the CFP, part of the interview involved showing students the census and the data entry software, and watching how they responded. These researchers found distinct differences in how prospective data entry assistants reacted to this primary source. Those who became very interested in the source, leaning forward and asking specific questions about it, were the ones hired. It is also important to gather references for prospective data entry operators, and not to hire anyone for whom you have no references. Investigators cannot just hire their favourite students unless they know those students have necessary experience with data entry. Before hiring, researchers should also review the guidelines offered by the Employment Standards Act for their province. For instance, the daughter of the one of the CFP principal investigators applied for a position. The B.C. Employment Standards Act showed that she was nevertheless eligible to apply, but her father then absented himself from the hiring process altogether so her hiring would be objective.

It is important to budget appropriate training time for hired staff and to provide them with a written user manual which includes everything from data entry procedures to rules on lunch breaks. Principal investigators must be familiar with university and provincial policies about sick leave and parental leave and how to deal with personnel problems or personality conflicts among the staff. Problems with programs or people need to be dealt with immediately before they damage morale or delay the project. As is the case in Minnesota, university employees may also be part of a union with its own employment standards which bear on data entry procedures. Researchers must also think in advance about what they will do if they discover a staff member cannot perform the work.

Contrary to their expectations that the arduous data entry must be part-time, the CFP found that its students were able to perform data entry full-time. The CFP research assistants entered data in eight-hour shifts over a three to four-month period with appropriate breaks. The Minnesota data entry operators also perform their work on a full-time basis, though Ruggles permits some flexibility in their work hours. One issue researchers will encounter in setting up data entry is whether to design a data entry program which is more restrictive or one which allows more discretion on the part of the data entry operator. Sager prefers the CFP approach in which data entry software was designed to minimize decision-making by the data entry operator. For example, the CFP data entry software featured a short list of acceptable entries for each variable which ensured data entry operators had an easy way to remind themselves what they could enter in each field. The CFP software also prevented data entry assistants from entering unacceptable information. Since the Minnesota Historical Census Projects has its data entry performed by professional staff members working year-round, some of whom have been with the project for ten years, it allows them more flexibility in entering data. For instance, the Minnesota data entry software features a DUPE or duplicator key, which allowed dwelling-level information to be duplicated from the previous accept. For individual-level variables, the Minnesota data entry operators could also duplicate entries from the person directly above and from related fields, such as parental birthplaces. The DUPE key is a risky feature, since it can encourage overly-rapid data entry resulting in mistakes. The CFP also permitted a DUPE key to be made part of its data entry program. In retrospect, however, Eric Sager advises that its use be severely limited. The Minnesota Projects also allows its data entry operators to make decisions about and sometimes correct faulty data. For example, if a Minnesota data entry operator encounters a household in which the enumerator wrongly recorded a widowed female head of household's relationship as "Widow" rather than "Head," the data entry operator can enter "Widow{HD}" in the field, preserving the original relationship and suggesting a correction. In contrast, the CFP investigators decided to emphasize accuracy over speed and convenience by insisting that its student data entry assistants enter exactly what they see on the census page. The CFP policy was that is was better to repeat mistakes in the original census document than allow other mistakes by giving the data entry assistant too much discretion. In both the Minnesota and CFP census projects, data entry operators could also enter comments in comment fields.

During the data entry phase of the project, it is important to have established a clear chain of command. The data entry staff must know who to ask questions when they arise, and who has final responsibility for decisions. Researchers must not leave their data entry assistants alone for long stretches of time, as questions invariably arise. It is also important to communicate decisions and solutions to other data entry operators, for example through a computer folder of problems such as odd entries and answers. Both the CFP and the Minnesota Census Projects had weekly meetings in which they discussed such problems and solutions. The data entry operators also need feedback on their work. Setting a minimum acceptable amount per week as the Minnesota projects do allows the DEOs to track their entry rates. Sager also advises that researchers conduct information sessions with data entry assistants to discuss the broader scope of the project, explaining why they are doing it in the first place. This helps the data entry assistants to understand the purpose of the final product.

Finally, project planners will have to decide whether the various members of the census project staff will conduct the stages of data entry, checking, cleaning, verifying and coding simultaneously, or one at a time. The CFP performed these stages one at a time. The data entry operators, research assistants and principal investigator of the Minnesota census projects have always attempted to perform their assorted functions of data entry, checking, cleaning and coding at once, though this simultaneity has become easier with the advent of more sophisticated spreadsheet software such as ACCESS for processing these data. Researchers must budget huge amounts of post-data entry time for verifying the data by conducting a re-entry of about 10% of the sampled data and comparing it to the first entry of that data.[18] Consistency checks must be made and errors fixed. If sample verification and checking is conducted simultaneously with data entry, the data entry operators should also receive feedback resulting from these activities. It is frequently necessary go back and change quite a bit of entered data. By Dr. Sager's estimation, the post-data entry stage takes about 50% of data entry time.

Dictionaries and Codebook

Finally, researchers must plan who will create the computer dictionaries in which data is translated to convenient numeric codes. They must also determine who will write the codebook which describes these data. At the CFP, these tasks were performed by Sager himself with the assistance of Doug Thompson and Marc Trottier. At the Minnesota Historical Census Projects, these jobs are performed by a group of twelve graduate student research assistants working out of one main room containing about eight computers. The Minnesota graduate assistants also write most of the codebook which describes their data, with Ruggles writing more complicated sections such as descriptions of sampling techniques and constructed variables. The programming of more sophisticated variables depends on the historian's own level of skill. At the Minnesota Historical Census Projects, the most sophisticated variables such as the constructed family relationship variables, were created by Ruggles himself. Ruggles picked up FORTRAN programming language during his own doctoral studies and uses this language for his more complicated coding needs. Marc Trottier has been creating versions of these same family relationship variables for the CFP using PARADOX.

Obviously most historians do not know FORTRAN; few even know SPSS. As a result, dictionary coding on most historical microdata projects will require close co-ordination with the hired computer consultant. Most of the needed historical variables, such as household- and dwelling-level characteristics, can be created in SPSS. In addition, once a programmer has written programs to create some of these variables, principal investigators can copy their syntax to create similar variables themselves. Since the historian is forever thinking of new questions, the research stage of historical microdata projects entails the endless creation of new variables. As a result, this learning process is important. It is crucial at this stage to ensure that the programmer and investigators keep careful records to document changes to existing variables and the construction of new ones. In this regard, the user-friendliness of the new, windows-based version of SPSS can work to the investigators' disadvantage. The most recent edition of SPSS software features a variety of windows and menu options which allow the user to make incremental changes to his or her database. Making these changes with successive clicks on the mouse is easy and convenient, but it also leaves the user without the "paper trail" of changes contained in the command files used in old Unix versions of SPSS. If investigators are dealing with a group of collaborating scholars, keeping a record of variable modifications which can then be shown to research partners upon request becomes even more important. Fortunately, the new version of SPSS continues to provide users with the option of to build command files as they work.

In previous publications, several historians have noted the importance of providing substantial documentation of historical microdata in a codebook. Gordon Darroch recommends that database documentation be written "in ordinary, nontechnical language to ensure a heritage of accessible nominal historical data." Secondary users must be able to determine fairly easily on the basis of this documentation whether they can use a historical database to explore their research questions. At the moment, the CFP documentation needs a more general introduction, which explains the research goals of its principal investigators and thereby sheds light on the wide range of historical questions which can be addressed through this source.[19] Nevertheless, the current CFP documentation is quite extensive, not only explaining the variables included in the data and listing their values, but also reprinting and explaining the enumerator instructions for the 1901 census, detailing the CFP sampling method and data entry procedures, and comparing totals based on the database to those in the original published census data. Most importantly, the CFP user's guide addresses the potential biases incorporated in the various census questions and the particular problems resulting from enumerator error.

The Minnesota Historical Census Projects have produced similar user's guides for each of its PUMS. The most important documentation innovation by the Minnesota project team has been to create a website which details their project. The principle aim of this website is to publicize and disseminate the IPUMS data. In the case of each PUMS, the Minnesota Historical Census Projects was granted funds to create microdata which were publicly available from the beginning rather than at the end of the project, as is the case with the CFP. As a result, the Minnesota project team has devoted greater effort to data dissemination than is currently appropriate for the CFP. The IPUMS website is quite extensive, allowing secondary users to download original copies of the IPUMS census data, as well as copies of the IPUMS documentation. Researchers who are only interested in one or two census years can download just the PUMS corresponding to those years, rather than the entire IPUMS. Other researchers who are only interested in specific variables, such as race and gender, can use an extract system to custom design a special sub-sample, or extract, of one or more census PUMS which includes only those variables. The IPUMS documentation is vast, comprising well over 500 pages. As a result, the Minnesota project team is now putting all its documentation on-line, including hyper-links to enable users to move from section to section. For example, the page, IPUMS Documentation, leads to information about the IPUMS Design, Variable Availability and specific information about each variable. The IPUMS website features a page for each variable, describing important information about what this variable meant in each different census year and listing the coding of values for that variable. Like the CFP documentation, the IPUMS website could benefit from a more general introduction which draws upon the Minnesota Census Projects' grant applications to discuss the broader historical applications of the IPUMS data. However, the design of this website, already constitutes a fine model for the documentation and dissemination of historical microdata.

Conclusion

Admittedly, these guidelines for creating historical microdata seem daunting. From forming a research collaborative to obtaining funding, then hiring a programmer, co-ordinating data entry and checking and cleaning data, to finally creating a codebook, the process of fashioning machine-readable historical data is full of challenges and potential problems--even setting aside the analysis, research and writing which is a historian's main business. However, the experiences of the CFP and the Minnesota Historical Census Projects show that there are many reasons for optimism. Project investigators should not be discouraged if their first attempt to fund a historical microdata project is unsuccessful. The CFP's first SSHRC application was turned down, while the Minnesota Census Projects' initial bid to create a new PUMS of the 1900 U.S. census also failed. In both cases, a second try met with success. Other challenges faced by microdata creators can be solved with money, time and advice from scholars who have previously created similar databases. In general, new software technologies are making the work of the principal investigator easier than ever. In Minnesota, for instance, the switch to ACCESS software for data entry and management is expected to make things easier for most project members. The consistent and constant feedback facilitated by the features of this software, which both manages and stores data, will allow experienced data entry operators to increase greatly the accuracy of the data. "A feeling of ownership by everyone involved," states Minnesota data entry operator Dianne Star, "increases the success of the project."

Historians with little previous experience in quantitative research should not think that creating microdata is an endeavour suited for others and not for themselves. Similarly, historians should not refrain from research employing routinely-generated resources because they think these sources are unusable. By bringing together interdisciplinary collaborators with a range of complementary skills, principal investigators can fashion projects in which partners' skill sets balance each other. Since many research questions overlap, it is possible to combine personal research interests with collaborative needs. The powerful insights suggested by layered analyses of a single source parallel the sort of complex knowledge which can result from studying historical microdata themselves. The strength of historical microdata inheres in the seemingly limitless combinations of variables which allow historians to investigate timeless historical questions. With both historical microdata and the academic collaboration which shapes and uses it, the outcome usually proves that the whole is greater than the sum of its parts. As a result, large historical microdata projects are worth the labour required to initiate and sustain them.

Notes

[1] The author would like to thank Eric Sager, Dianne Star, Todd Gardner and Peter Baskerville for their comments on drafts of this paper.

[2] Steven Ruggles and Russell R. Menard, "The Minnesota Historical Census Projects," Historical Methods: A Journal of Quantitative and Interdisciplinary History, The Minnesota Historical Census Projects (Special Issue), Vol. 28, No. 1 (Winter 1995), 6.

[3] See Kris Inwood and Richard Reid, "Introduction: The Use of Census Manuscript Data for Historical Research," Histoire Sociale/Social History, Vol. 56 (November 1995): 301-311 for more detailed information on the range of historical microdata available in Canada.

[4] Historical Methods: A Journal of Quantitative and Interdisciplinary History, The Minnesota Historical Census Projects (Special Issue), Vol. 28, No. 1 (Winter 1995); Matthew Woollard, "The Nineteenth-Century Censuses Collection," Paper presented to the Social Science History Association Conference, Washington, D.C., 1997; "Historisk Institutt, Bergen" website, http://www.uib.no/hi/1801page.html. See the website, "Historical Microdata Around the World," http://www.isv.uit.no/seksjon/rhd/nhdc/micro.htm, for further information about international historical microdata.

[5] See, for example, Inwood and Reid; Jose E. Igartua, Gérard Bouchard, Hubert Charbonneau, David Gagan, Gordon Darroch and Chad Gaffield, "Table ronde--Round Table: Les bases de données historiques: L'expérience canadienne depuis quinze ans/Historical Database: The Canadian Experience Since Fifteen Years," Histoire Sociale/Social History, Vol. 21, No. 42 (November 1988): 283-317; David P. Gagan, "Enumerator's Instructions for the Census of Canada, 1852 and 1861," Histoire Sociale/Social History, Vol. 7, No. 14 (November 1974): 355-365; J. Dennis Willigan and Katherine A. Lynch, "Chapter 4: Enumerations and Censuses," Sources and Methods of Historical Demography, (New York: Academic Press): 79-108; Ian Winchester, "Review of Peel County History Project and the Saguenay Project," Histoire Sociale/Social History, Vol. 13, No. 25 (May 1980): 195-205; D.I. Pool, "The Historiographer in Computerland: A Review Article," Histoire Sociale/Social History, Vol. 8, No. 15 (May 1975): 165-174; Chad Gaffield, "Theory & Method in Canadian Historical Demography," Archivaria, No. 14 (Summer 1982): 123-136; Kris Inwood, "The Representation of Industry in the Canadian Census, 1871-1891," Histoire Sociale/Social History, Vol. 56 (November 1995): 347-373; Peter Baskerville and Eric Sager, "Research Note: Finding the Work Force in the 1901 Census of Canada," Histoire Sociale/Social History, Vol. 56 (November 1995): 521-535; Nancy Folbre and Marjorie Abel, "Women's Work and Women's Households: Gender Bias in the U.S. Census," Social Research, Vol. 56, No. 3 (Autumn 1989): 545-569; Alan A. Brookes, "'Doing the Best I Can': The Taking of the 1861 New Brunswick Census," Histoire Sociale/Social History, Vol. 9, No. 17 (May 1976): 70-91; Steven Ruggles, "Historical Demography from the Census: Applications of the American Census Microdata Files," in David S. Reher and Roger Schofield, eds., Old and New Methods in Historical Demography, (Oxford: Clarendon Press, 1993): 383-393; Lutz K. Berkner, "The Use and Misuse of Census Data for the Historical Analysis of Family Structure," Journal of Interdisciplinary History, Vol. 5, No. 4 (Spring 1975): 721-738.

6. Ruggles and Menard; Gordon Darroch, "A Study of Census Manuscript Data for Central Ontario, 1861-1871: Reflections on a Project and on Historical Archives," Histoire Sociale/Social History, Vol. 21, No. 42 (November 1988): 304-311; R. Christian Johnson, "A Procedure for Sampling the Manuscript Census Schedules," Journal of Interdisciplinary History, Vol. 8, No. 3 (Winter 1978): 515-530; Michael D. Ornstein, "Discrete Multivariate Analysis: An Example from the 1871 Canadian Census," Historical Methods, Vol. 16, No. 3 (Summer 1983): 101-108; Gérard Bouchard, "Current Issues and New Prospects for Computerized Record Linkage in the Province of Québec," Historical Methods, Vol. 25, No. 2 (Spring 1992): 67-73; Elizabeth Bloomfield, "Research Note: Using the 1871 Census Manuscript Industrial Schedules: A Machine-Readable Source for Social Historians," Histoire Sociale/Social History, Vol. 19, No. 38 (November 1986): 427-441; Michael J. Doucet, "Discriminant Analysis and the Delineation of Household Structure: Toward a Solution to the Boarder/Relative Problem on the 1871 Canadian Census," Historical Methods Newsletter, Vol. 10, No. 4 (Fall, 1977): 149-157; Lisa Y. Dillon, "Integrating Nineteenth-Century Canadian and American Census Data Sets," Computers and the Humanities, Vol. 30 (1997): 381-392.

7. Ruggles and Menard, 6-8.

8. The 1890 U.S. manuscript census burned in a fire, preventing the possibility of creating a PUMS for that year. Ruggles and Menard, 7.

9. David Gagan, "Some Comments on the Canadian Experience with Historical Databases," Histoire Sociale/Social History, Vol. 21, No. 42 (November 1988): 303.

10. Eric Sager et. al., "Description of Proposed Research: The Canadian Families Project," Application for a Multiple Collaborative Research Initiative Grant, Social Sciences and Humanities Research Council, June, 1995.

11. Gaffield, "Theory and Method…," 135.

12. Here, the CFP requested proportionately more than the Minnesota Projects to facilitate the research phase of their project.

13. In its proposals, the Minnesota Historical Census Projects allocated its research assistants to these functions: budget officer, personnel co-ordinator, co-ordinator of documentation and procedural historian, data entry co-ordinator and data cleaning and quality co-ordinator. If data entry operators are public service employees, it may not be permissible to have a student supervise them; another supervisor, such as a principal investigator, may have to be appointed in accordance with policy.

14. Project investigators in both Minnesota and Victoria recommend purchasing these microfilms rather than borrowing them on inter-library loan, as checkers and cleaners will need to re-order films for cases which must be altered.

[15] Steven Ruggles, "Public Use Microdata Samples of the 1860 & 1870 Censuses," Grant Application to the Department of Health and Human Services Public Health Service, September 27, 1995, 2.

[16] Darroch, 304.

[17] Darroch, 305.

[18] For 1860/1870 PUMS, the Minnesota census projects is verifying one out of 25 reels.

[19] Eric Sager, Douglas K. Thompson and Marc Trottier, The National Sample of the 1901 Census of Canada: User's Guide, Version 1.0, (Victoria: Public History Research Group, University of Victoria, 1997).