Canadian Century Research Infrastructure Project:
People and Place in Historical Microdata

A Paper Completed as Background Research for the Geography Unit of the CCRI

Rev. 5

Justin J. Ngan
Byron Moldofsky

Department of Geography Program in Planning
University of Toronto 2004
1. Introduction

Historical microdata projects have been undertaken by various countries including
Canada. Microdata, or individual level responses recorded from census enumerations,
offer researchers the ability to ask questions about people, families, and society that
cannot be answered through examination of aggregate census statistics alone. The
meaning of aggregate statistics can change from census to census because of the
redefinition of measures at different points in time and because of the shifting importance
placed on various measures or indices. In asking questions to understand human
processes, their relationships, and their consequences, researchers can use historical
microdata to create new variables and indices of interest that were not of any concern in
all or some of the past censuses.

Researchers interested in understanding the people and places of the past and present,
look to many sources such as the manuscript enumeration forms from which microdata is
compiled. This source is favored for its accuracy and continuity. Despite problems of
rectifying contextual questions and issues of consistency between census years, it is a rich
source of quality data compared to most other available historical data sources. At the
same time however, while this data is a valued archive of information, it remains
inaccessible to most researchers because of the technical expertise required, the immense
cost to assemble and transcribe the information from paper manuscripts and, the
requirement to keep enumeration records anonymized. A historical census microdata
project that produces anonymized records for research purposes is therefore valuable for
its cost efficiency and for the information that it makes available for constructing
knowledge of patterns, changes, and outcomes of people, past and present.

The purpose of this paper is to offer a geographic perspective of the creation and use of
historical microdata as the Canadian Century Research Infrastructure embarks on its
mission to construct microdata samples for the Canadian Census years 1911 through
1951. While the project is in its early stages, considerations at this time are important in
steering the project to meet best the needs of the wider academic community. The paper
is divided into several sections beginning with a look at some of the past microdata
projects that have been undertaken. The CCRI benefits from the experience gained and

---

1 This list of microdata projects includes: the Individual Public Use Microdata Samples (IPUMS) project
undertaken at the University of Minnesota, notable for being the most comprehensive single effort, the
Canadian Families Project (CFP) which created microdata for the 1901 Census year, the History Data
Service in the UK, the Norwegian Historical Data Centre, Sweden’s Demographic Database, The Danish
Demographic Database, and the Historical Sample of the Netherlands. See Appendix 1.

2 In particular, ‘modern censuses’ are valuable for the four qualities that make them ‘modern’: they are
universal encompassing the entire population of a country, they are simultaneously collected, they are
census/census6.htm.

3 Anonymity guidelines vary among different countries and issues regarding the guarantee of privacy for
respondents examine in detail. See http://www12.statcan.ca/english/census01/Info/finalrep.cfm#2.8%
20The%2092-Year%20Rule for a detailed report issued by the appointed expert panel for recommendations
on handling Canada’s historical Census records.
the expertise developed from similar endeavours. Thus a review of the past projects will both recognize past contributions and offer an understanding of the different nature of various historical microdata projects. After a brief review of past projects, the importance of microdata as a data source is considered. In particular, a perspective of historical microdata in geographic context is presented to offer some reasons of why a consideration of space is quite pertinent in the project’s endeavour. Given such a geographic perspective, the paper presents two primary considerations that may warrant further investigation at this time.

### 2. Historical Microdata Projects

#### 2.0 Introduction: National Historical Microdata

In most countries, an assortment of recorded data about its people exists in various forms that may include in addition to census manuscripts: population registers, business directories, phone books, obituaries, birth records, tax records, and immigration records (see Appendix 1 for comparison of some selected projects). The importance of compiling this data may primarily be to offer a rich source of accessible information that can be used to examine the past, but it is also to ensure the preservation of that information. Otherwise complete time-series can become fragmented as a result of various events such as the fire that destroyed the manuscript records of the 1890 United States Census manuscripts.

The practice of census enumeration has varied between different countries although attempts are underway at the United Nations, Statistics Division to bring the expertise of different countries together for the purpose of sharing strategies and methods. The historical microdata vary between different countries because of the differences in what censuses recorded and how people were enumerated. Similarities exist between censuses of North American countries primarily because Canada used the United States Census as one of its models and both countries, compared to those of Europe, are relatively young with modern censuses covering most of their census recording years.

Few countries have re-constructed historical microdata with the longitudinal coverage that is offered by the United States. Even when attempting to compare short periods, perhaps of single census periods, the differences between countries can be sufficient to pose challenges for even an experienced researcher let alone one with limited resources. It has been found that comparability of harmonized microdata is helpful (Dillon 1996; Dillon 2000; Dillon and Thorvaldsen 2000). Exploring challenges of harmonization have begun, now that a greater number of digitized historical microdata have been produced. I refer the reader to the “Handbook of International Historical Microdata for Population Research” for a more detailed description of each country’s dataset and contributions. Highlights from the Handbook follow to provide a flavour of the nature of some past projects.
2.1 Argentina

It is apparent from the section in the handbook covering Argentina’s historical microdata project, that it is considered an influential piece of pioneering work. Despite the project’s release in 1967, well before the concept of metadata and data dictionaries were popularized by increased use of geographic information systems, the database produced was well documented, which is considered a crucial component of any distributed data today. Given the level of programming expertise that would have been required in the 1960’s to construct a digital database, the project likely benefited from good programming practices that would come to underlie the computer programming discipline. The project is also recognized for its accessibility. It was provided freely to researchers allowing for the extraction of value through the construction of knowledge about the past. This arguably is the goal of any such endeavour. Technical limitations of the time have made the database susceptible to loss in value. The loss of the original digital tapes and punch cards has resulted in the database losing its value because over time and repeated transmission, the records have been incorrectly sorted. The project remains a monument to the achievement of its creators and their vision in making the dataset freely available.

2.2 Great Britain

Among Great Britain’s most used historical data are the microdata samples for the 1851 and 1881 Censuses (University of Essex 2003). These two microdata samples form the core of the Great Britain Historical Database and the Great Britain Historical GIS Programme. The historical database is supplemented by a rich collection of socio-economic statistics for the 19th and 20th centuries. While the data collection does not have the longitudinal continuity of the U.S. collection, one of the noteworthy aspects of the Great Britain historical data is the impressive use of a Geographic Information System to create a geographic database capable of mapping available statistics with accurate boundary definitions for any given date. Such utility is indispensable given the boundary changes over time. The Great Britain Historical GIS Programme has recognized at least two reasons for the importance of spatial frameworks when working with historical microdata including: giving meaning to historical place names that may otherwise be meaningless and “studying the fine geographical structure of past societies” (Gregory and Southall 2003).

2.3 Denmark

The creation of historical microdata in Denmark has relied on informal collaborative efforts of amateur and professional historians as a result of a lack in public funding to support a large scale reconstruction project (Clausen and Marker 2000). Nevertheless, an impressive structure has been created to administer the collection and dissemination of historical microdata contributions from smaller scale efforts. This has come from the development of research data archival facilities over the last quarter of the twentieth-century and from the effort and direction of the Cooperation Committee for Source Entries (SAKI). SAKI has probably come to life as a response to the need for

---

4 FORTRAN, the first high level programming language was introduced in 1954 long before a metadata standard was introduced in the United States by a bill signed by President Bill Clinton in 1994.
standardizing the creation of digital datasets in the absence of the funding needed to undertake a single and more comprehensive effort.

There are clear barriers facing those who wish to explore and document the changes in Denmark’s history with greater spatial and longitudinal coverage than is currently available. Forced to handle analysis constrained by the costs of digitizing datasets, researchers are contributing in piecemeal fashion to the creation of a historical microdata collection with gaps being filled over time. The freedom of researchers to choose a data source from options that include (in addition to censuses) parish registers, land registers, and conscription registers means that the database has the potential to become a source of historical information that is quite rich, assuming the preserved historical sources are not lost from unforeseen circumstances. To this end, the Danish Data Archive (DDA) administers the access to and digitization of historical data to avoid duplication of effort and maintain quality standards.

2.4 Netherlands

The historical microdata of the Netherlands are unique in its design. Specifically, while most historical microdata have relied on census data for part or its entire database, the Historical Sample of the Netherlands (HSN) relies on birth certificates as the source for the sample. It is recognized that the approach excludes immigrants. One reason that this approach may have been chosen is that the population registers were administered by some 1300 municipalities around the middle of the 19th century leading to problems of rectifying inconsistencies as well as missing data due to loss from events such as war (Mandemakers 2000). To account for exclusion of foreign born residents, intentions have been expressed to create a separate sample from census and population registers.

2.5 United States

The American historical microdata represents the single largest historical microdata construction effort to-date, spanning the censuses from 1850-2000. The Integrated Public Use Microdata Series (IPUMS) is noteworthy on several accounts including: longitudinal continuity, longitudinal comparability, documentation, and accessibility. Following the model for release of its decennial Census data, the IPUMS data is available for free through its online extraction system to the general public.5 The commitment of both the U.S. government and the U.S. Census Bureau in ensuring the free distribution of its censuses’ aggregate and microdata is unique. The samples that comprise the IPUMS series span 1850-2000. While the tight integration of the IPUMS series would lead one to believe that the series was the result of a single planned effort, it is in fact the culmination of various groups at various times. A brief review of the timeline suggests that the earliest interest in U.S. historical microdata began with Stephen Thernstrom’s use in 1963.

At the same time, the U.S. Census also released the first national census microdata for the 1960 Census. With the release of this and the 1970 microdata series, two efforts were undertaken to extend back in time the historical microdata series. Samuel Preston at the

5 http://www.ipums.org/cgi-bin/usa/extract_main.pl?process=IntroLogon
University of Washington and the University of Pennsylvania produced a 1 in 750 sample of the 1900 census and a 1 in 250 sample of the 1910 census. Halliman Winsborough at the University of Wisconsin in collaboration with the Census Bureau created 1 in 100 samples for the census years 1940 and 1950. Ruggles and Menard continued the effort to reconstruct the past microdata series with work on the 1880 Census with funding from the National Institutes of Health. They then continued with the 1850 Census. The 1850 Census year marks the start of the IPUMS series because it was the first year that the census began to collect information at the level of individuals.

The strength in the IPUMS series is not only in its continuity, but also in its comparability. As part of the reconstruction project, efforts to recode answers and harmonize responses between census years has made the job of making comparisons across census years less cumbersome for researchers. Details regarding sampling strategies, recoding and other contextual questions are all readily available online. Documentation has contributed strongly to preserving the value of the microdata series. While the online extraction system makes the data series physically accessible, the documentation has made the dataset technically accessible.

In 2002, construction began on The National Historical Geographic Information System. It is being built to provide a geographic database of aggregate census data from 1790 to 2000, and as a complement to the IPUMS project. It is yet another major endeavour of the United States to make available a research infrastructure for studying the people and places of its past while overcoming the challenges of reconciling changes in census geography (Fitch and Ruggles 2003).

2.6 Canada

The Canadian Families Project (CFP) is a precursor to the current Canada Century Research Infrastructure (CCRI). It involved the creation of digital microdata for the 1901 Census as an initial project to assess the methods and gauge the value of a Canadian historical microdata creation effort. The sampling framework of the project was designed to simultaneously provide the researcher with the information necessary for examination of both individuals and the household units to which they belonged (Baskerville 2000).

In addition to an increased understanding of the social structure of Canadian families in the early twentieth-century, the project has demonstrated the value that can be achieved from understanding the past, such as proving or dispelling assumptions of Canadian society, particularly those that may have constructed correctly or incorrectly, notions of what made Canada and its people and structures unique. Baskerville (2000) identified a challenge inherent in the creation of the 1901 Census data as being the quality of the spatial analytical framework. Specifically, the lack of reference maps to the enumeration boundaries or polling subdivisions that were used in the manuscripts has resulted in a limitation of the degree to which spatial variability at scales finer than large census sub-districts can be ascertained.

---

6 Baskerville (2000) cites the discrepancy in counts of divorced individuals. He notes that the source of the discrepancy is likely the re-coding of responses by the Census office, and the result has been an incorrect assumption of the social reproduction of families in early twentieth-century Canada.
The reconstruction of a Geographical Framework for the 1901 Census of Canada foreshadowed many of the challenges that are now faced by the CCRI team. The purpose of reconstructing the geographic framework remains similar: “to interpret the spatial processes, patterns, and structures associated with family life in Canada’s cities and rural areas” (Buck, Jordan et al. 2000). The challenges facing the CCRI and eventually researchers working with the data appear to be unfolding in a similar manner. These include difficulties such as a lack of comprehensive cartographic documents identifying statistical boundaries, the fluidity of their boundary definitions during a period of rapid expansion, and issues of varying areal unit sizes, discussed in more detail later in this paper.

The CCRI can further the analysis and understanding of Canada’s past and its influence on the present with a database of historical microdata. Part of the CCRI’s work is to explore the issues associated with spatial statistics, including those that may limit the accuracy and interpretation of microdata and aggregate results. This will improve the resolution and clarity of the lens that can be used to examine the past and as an example, add to the understanding of how the reproduction of Canadian society has occurred throughout the country.

2.7 Summary of Historical Microdata Projects

The importance of historical microdata as both a way to preserve historical information and as a means to help researchers understand the past is evident in the variety of ways that historical microdata has been collected and used. A difference is noticeable between the source and structure of microdata for the European countries and those of the new countries of Canada and the United States.

The European countries, with longer histories and local customs have benefited in some ways and suffered in others when it comes to having a reliable source of data from which to build their historical microdata. On the other hand, Canada and the United States have geographies that have remained relatively stable over the 19th and 20th centuries, notwithstanding the shifting internal boundaries of rapid westward expansion experienced in both countries during this period. Both countries have also benefited from a relatively uniform method by which census enumeration has occurred with most of the period of interest covered by so-called modern-day censuses. It should also be recognized that few countries have the benefit of extensive funding to complete such projects like those carried out in the United States and Canada, yet the efforts of committed researchers have made it possible to produce formidable databases.

3. Microdata

3.0 Structure and Context

Microdata are datasets of enumeration level responses to census questionnaires or other individual level responses to surveys. They vary in how the collections of responses are structured, sampled, and spatially referenced. Hierarchical structures such as those
employed in the construction of the Canadian Families Project allow a researcher to examine the microdata at different interrelated scales. Researchers can examine the data by looking at characteristics of individuals. Alternatively, they can examine the characteristics of families as formed by individuals. They can also focus on households and uncover the patterns of social relationships that are common at particular periods and how they have changed over time. The sampling design employed to construct microdata is an art in balancing various factors including: achieving a ‘desirable’ sampling ratio, retaining embedded levels of information through structural hierarchies, maximizing the value of the effort by uncovering the most about our past society at the least generalized scale possible, and doing all this within cost constraints.

The sampling ratio or sampling density, determines how representative the samples are of the population. The sampling density can vary not just from one dataset to the next, but also within a particular dataset. An unweighted ‘flat’ sample has a single sampling density used to derive all records in the sample; whereas a weighted sample is one in which records of persons with particular characteristics are intentionally over- or under-represented. Over-representation might be intentional so that accuracy and precision can be increased. Sampling densities can therefore increase the knowledge derived from the sample, but only through the cost incurred in design and construction.

The spatial reference or geographic framework of the microdata, determines the degree to which a researcher is able to make insightful observations and informed conclusions about the population in a geographic context – of telling the most that one can about the population at the least generalized scale that is possible. A sampling scheme enables researchers to discern informative patterns to recognize regional and local patterns and differences. Geographic considerations occur early in the design of datasets when decisions are made regarding how to capture a representative sample of a population that is not randomly distributed throughout the country. Populations tend to be clustered leading to varying population densities. Populations also tend to organize into spatial patterns of distribution according to various measures of socio-economic variables including, race, income, and age. Recognizing these patterns at the onset can improve the quality of the dataset that is created. Geographic considerations also arise when it comes to how the data is to be anonymized, accessed, and distributed. Inclusion of a substantive geographic framework with microdata can help researchers in identifying and analyzing spatial patterns.

These considerations become even more critical when approaching longitudinal analyses to identify spatial-temporal shifts. As the authors indicate in the article introducing the IPUMS Redesign (Ruggles, Sobek et al. 2003), “inconsistencies in the coding of geographic areas have long posed frustrating compatibility issues for public microdata users.” Microdata offer researchers a unique opportunity to gain an “historical perspective” of the socio-economic characteristics and processes occurring through time and space, without the limitations imposed by aggregate level data reporting.

It is important to recognize the contribution of historians and sociologists to the “contextual component” of the project. Without the knowledge and understanding which
they bring and the work they undertake, any subsequent effort to make sense of the data and the people over the many decades would be stifled by the various contextual questions that arise. More specifically, their contribution in the construction and interpretation of microdata is to provide the context in which census questions were asked and understood. As described by Sager (2000), censuses are “surveys reporting the voices of large numbers of people…[the] results should be understood as a dialogue, a long series of questions and answers in which class, race, gender, language, and other influences guide the conversations.” Understanding the context of the dialogue enables the researcher to make sensible findings from any one particular census or across dialogues that span numerous decades.

3.1 Microdata v. Aggregate Data

Aggregate data reported for censuses are convenient and useful. They offer generalizations of popular socio-economic variables and indices and make it convenient to understand characteristics of the population at-large for even relatively small areas such as the census subdivision, tract, or block.

Some aggregate statistics reported for censuses are arguably interpretations *de jour*. Reported aggregate statistics may be calculated variables, which incorporate classifications that are deemed to be relevant for the period in which they were reported. The value placed on some particular calculations or classifications may be more ephemeral or contestable than others. We see this in the case of the low-income cutoff (LICO) statistic that has been reported as part of the Canadian censuses of the recent decades.

One of the difficulties associated with the use of aggregate census data for longitudinal studies is the continuous redefinition and reinterpretation of aggregate variables and indices. Redefinition of aggregate census variables is evidenced by the census dictionaries that have accompanied census releases. They detail differences in how variables are defined and calculated. Reinterpretation can occur even without official intervention by statisticians. Over time, changes in social perceptions may imbue variables with different meanings. Even for cases in which census-to-census changes can be rectified, an uncertain level of error can be introduced in the process and the implications for interpreting the results can complicate the task of the researcher. A researcher, looking at the past through the lens of contemporary thought, might find that the aggregate measure of interest only has been reported for the last couple of censuses.

More likely is the case that a researcher faces the inadequacy of aggregate data to provide the information about people that they require. That is, they need to understand the structure and characteristics of society as measured in such a way that is not reported through aggregate data. It is simple to ask a question such as how did the age structure of the population vary throughout rural areas of Canada in 1996 and find the answer in aggregate census data, even if we overlook the fact that what is considered ‘rural’ has changed over time. It is more difficult to find an answer to a question such as how has social reproduction changed throughout the twentieth-century as measured by what activities people are engaged in during different periods of their lives as classified by age.
While many would agree to or accept the defacto categories reported in the aggregate census tabulations, there would be much less understanding and perhaps agreement on the definitions required to construct a measure for the second question. Microdata offer historians and other social scientists an opportunity to engage in discourse on issues from a historical perspective. It offers the researcher the ability to examine societies of the past from contemporary perspectives; uncovering ideas about who we are that would otherwise have remained hidden by past assumptions.

3.2 Microdata and Geography

3.2.1 Data Modeling in Geography

Geographic databases differ from those used in general statistical analysis because of the implicit incorporation of location as an attribute. Location becomes an important attribute because it is considered that results of analyses are dependent on where things are in space (Goodchild 2001). Geographic databases are based in part on how the geographic world is modeled. For anthropogenic features and processes, the model used is often discrete. Discrete models differ from continuous models in that entities in the model are located in specific points in space. Continuous models are often more appropriate in representing natural features and processes such as land elevation, bathymetry or air quality (Burrough and Donnell 1998; DeMers 2000).

In the case of constructing a geographic framework for historical microdata, what is in fact being created is a geographic model of the world, in which people are entities with particular attributes, distributed throughout a finite space partitioned by census and political boundaries. The attributes associated with each entity is the rich information provided from responses in census enumeration manuscripts. Such a model requires that researchers have a grasp of the underlying paradigms of spatial interaction and the intricacies involved in spatial statistics.

Two concepts in Geography seem particularly relevant in the construction of a geographic framework for historical microdata, or the spatial modeling of people and place in history: Tobler’s first law of geography (Tobler 1970) and the modifiable areal unit problem (Openshaw 1977). Tobler’s first law states that everything is related to everything else, and that things closer together tend to be more alike than things further apart. Certainly, if Tobler’s first law of geography is indeed true, then one can see some of the implications for constructing microdata including the need to understand how people are in fact related to each other over space and to understand the correlation and dependence between the variables that are recorded.

The modifiable areal unit problem addresses the issue of differences in the results of analysis depending on the areal unit used. Two sets of effects are pertinent to this discussion: scale effects, and aggregation or zone effects. These effects limit or enhance the degree to which spatial patterns can be substantiated. It has already been acknowledged that the construction of microdata involves cost constraints. Given these constraints and the need to ensure anonymity of respondents, we are confronted with the questions: at what scale do we sample the population, what geographic scale is most
appropriate for the identification of sampled records and, what is the effect of aggregating individual level responses? Further, we find that as decisions of scale are made, decisions of aggregation or zonation may arise. The works of Openshaw (1977; 1978), Fotheringham and Wong (1991), and Amrhein (1995) illustrate the complexities involved with the modifiable areal unit problem. While some statistics are robust and withstand the effect of the MAUP, others are less so. For others still, their resilience seems to depend on the underlying variance in the population.

3.2.2 Tobler’s First Law of Geography

Tobler’s first law of geography, embodying models of spatial structure including the Central Place models of Christaller and Lösch and the diffusion models of Hägerstrand, provides the underlying assumption for spatial interaction (Johnston, Gregory et al. 2000). That assumption is what I suggest to be the reason for the importance of Geography’s role in the construction of historical microdata. It offers a basis for considering the methodology for a spatial sampling framework, bringing to light opportunities from spatial interaction that can be used advantageously as well as recognizing difficulties that may be encountered. For example, we consider that part of the importance of constructing the microdata for the 20th century is that it was a period of rapid settlement and urbanization in Canada. Bringing what we understand of ideas such as chain migration to mind, one might consider the possibility that people organized themselves very early in a non-random fashion based on their country of origin, occupation, or period of immigration.

In fact, we already recognize that this is the case, and we see the heritage of those early 20th century settlers as evidenced by local place names like Cabbagetown and the Holland Marsh. Understanding the creation of such social spaces (Buttimer 1969) and those less obvious that have yet to be uncovered from studying the past is the opportunity that microdata affords us. This is also the context of which we must be creatively aware when constructing a spatial sampling framework.

A spatial sampling framework for building microdata necessarily takes on the knowledge and imagination of its principal investigators. It does so to keep open the possible avenues of investigation that can lead to constructing knowledge about the past. In particular, it encompasses the ideas of likely scenarios and important influences in order that information can be drawn from the data with sound methodology. For example, did new immigrants live with relatives who were already settled when they first arrived? How long did new immigrants maintain household relationships with relatives? Were settlers of the expanding West different from those who settled in more established regions?
Understanding the likely scenarios and important influences that come from spatial interaction therefore means that the relationships between variables over space can be accounted for *a priori* in order to decide the necessary sampling densities and spatial resolutions required for particular regions and locations. It also means anticipating later difficulties that researchers may face such as high cross-correlations between variables or the need to aggregate individual responses.

We see this in the approach used in constructing the American microdata by the IPUMS group. To understand the influence of the Civil War on the migration of blacks, blacks were over-sampled at twice the density of other households for the 1860 and 1870 censuses (Ruggles, Sobek et al. 2003). In the exploration of the past, we are not always privileged to know in advance such important influences and interactions. After all, part of the exploration is indeed to uncover such relationships and processes and prove or disprove what we know. Given this scenario, a strategy that relies on maximizing the usefulness of samples based on population distributions and densities becomes a possible way to preserve the embedded information. The role of a geographic framework then becomes even more important not only in constructing the sample, but in assisting the researcher to make sense of their findings.

As part of the IPUMS Redesign, it is acknowledged that a lack of a consistent geographic framework has been a problem for researchers (Ruggles, Sobek et al. 2003). The need for researchers to understand how consistently defined areas of an interesting spatial resolution have changed with regard to constructed measures, is necessary if the researcher is to draw conclusions of significance for anything other than region-wide or even country-wide contexts. A suitably constructed geographic framework can unlock potential for contextual interpretation that may interest researchers in formulating and answering questions about our past.
Map Plate 1 demonstrates how an increasing resolution in the geographic framework can improve the sampling accuracy, the accuracy of sample point representation, the ability to discern patterns in the sample, and the ability to improve interpretation of patterns from knowledge of local context. On Map A, a lack of partitions for the depicted area means that sample points must be chosen from the area as a whole. Without a geographic framework stratifying the area, sample points will be chosen randomly from the whole area. This may mean that areas of greater densities may be under-represented. If Wirth’s (1938) idea that one characteristic differentiating urban and rural areas is an increase in heterogeneity in urban locations, then it may be possible that the sample will not adequately capture such differences when insufficient sample points are drawn from the more densely populated urban locations.

The accuracy of locating sample points is also the lowest on Map A (Plate 1). A researcher only has knowledge that a sample point lies somewhere in the general area, whereas on each subsequent map, B through E, accuracy of sample point location improves. The importance of this increasing accuracy is evident as a population distribution pattern begins to emerge progressively through the maps. Note how the increasing number of partitions in the upper part of the peninsula, sorts out the location of sample points and reveals a cluster pattern less accurately depicted on maps B and C and not shown at all on map A.

Given that a researcher has been able to illustrate the pattern shown in map (e), the geographic framework can be further used to interpret the results. Context in the form of local characteristics can then be used to help provide a rationale for the observed pattern. For example, the Public Use Microdata Area (PUMA) in the western part of the map is sparsely populated. Local geography can help to explain this pattern. Alternatively, PUMAs in the head of the peninsula do not show the same densities. Knowledge of factors such as house types, prices, and ethnic distributions may help to explain the differences.

Map Plates 2 through 5 advance the notion that an increasing resolution for the geographic framework can yield important information for both the construction of the microdata and its subsequent use and interpretation. Each plate shows the distribution of an ethnic group as reported in the 2000 U.S. Census microdata for two classes of education attainment. Just as in Map Plate 1, the increasing accuracy of sample point locations and the ability to depict distribution patterns is evident. In addition to this, it can be seen that within any particular group, variations in the distribution does not only apply to the location of sample points, but also to characteristics of the sample points. Persons of different levels of education attainment appear to have spatial preference and this preference is not the same across the different ethnic groups. Such differences can be important when considering alternative sampling strategies and when examining differences in the population such as social behaviour.

Also notice that the patterns that emerge on one map do not necessarily hold true in a subsequent map with finer partitions. Take the eastern side of the map just below the
head of the peninsula. The degree of segregation is quite different throughout the series of maps and quite a different story might be told based on each.
Map Plate 1. Increasing Geographic Resolution: Accuracy, Pattern, Context

This series of maps shows an increasing level of geographic resolution with Map E. showing the 5% Public Use Microdata Areas of San Francisco and part of the South Bay area.

The result of an increase in the geographic resolution is an increase in the accuracy of depicting the population distribution, an increase in the ability to discern the pattern in distribution, and an increased ability to place that distribution in relevant context, in this case, areas defined as urban in Map E.

Justin Ngan, 2004

1 dot = 2,500 persons

Source: U.S. Census Bureau, 2000
Map Plate 2. Distribution by Education Attainment in the San Francisco & the South Bay: White respondents

This map plate shows differences in the distribution of respondents by education attainment. This series of maps demonstrates the benefit of increasing the spatial resolution of areal units, with implications for sampling considerations when producing microdata samples and for research findings.

Map A has no spatial partitions. As a result, it is not possible to identify patterns in location of respondents by education attainment. Map B has large, coarse partitions, as a result, patterns in density and segregation start to emerge. The top left aggregation of PUMAs seems to be denser than the others. In addition, the aggregation in the bottom right appears to show an emerging concentration of those with post-graduate education attainment. Map C refines the accuracy of the density distribution, identifying the nature of the more sparsely settled PUMA to the southwest. In addition, Map C and Map also Map D, shows the segregation pattern of respondents by education attainment more prominently. In Map E, the 5% PUMAs are shown and the concentration of respondents with post-graduate education attainment is noticeable in the north and southeast areas.
Map Plate 3. Distribution by Education Attainment in the San Francisco & the South Bay:
Hispanic respondents

This map plate shows the distribution of Hispanic respondents by education attainment and the benefits of increasing spatial resolution as in the previous plate.

In comparison to that plate, the first noticeable difference is the far fewer number of those with post-graduate attainment. In addition, the prominent clustering in the east side of the peninsula stands in contrast to the lower density in the northern PUMAs of the peninsula. In the southeast, the higher densities shown in the map plate of white respondents are not present on this plate.
Map Plate 4. Distribution by Education Attainment in the San Francisco & the South Bay:

Black respondents

In comparison to the other plates, there is a noticeable clustering of Black respondents in the east side of the peninsula's head. There are far fewer Black respondents in the other PUMAs relative to White and Hispanic respondents.
Map Plate 5. Distribution by Education Attainment in the San Francisco & the South Bay: American Indian respondents

The American Indian respondent is the least densely distributed. Similar patterns of distribution to the other groups are present, such as the concentrations in the head of the peninsula. This concentration may be related to the higher population densities in the area close to San Francisco.
To summarize, Tobler’s first law of geography offers one perspective of the need to consider and incorporate a spatial framework in the microdata construction process, which can be used subsequently in the analyses performed by researchers. Not only is a spatial framework necessary, but one of a reasonably fine resolution can contribute to the value of the microdata and the richness of the findings from research. Spatial partitioning offers the ability to interpret results of analyses within specific regional and local contexts, which is important in providing insight into processes and patterns of spatial interaction and spatial organization. As the fineness of that partitioning increases, so too does the accuracy of the computed measures, increasing the likelihood of identifying or confirming patterns and relationships of socio-economic behaviour.

3.2.3 The Modifiable Areal Unit Problem

The modifiable areal unit problem (MAUP) has posed challenges for social scientists engaged in research that involve the aggregation of data at the individual level to levels suitable for analysis and anonymized reporting. Specifically, two effects are associated with the modifiable areal unit problem, scale and aggregation (Openshaw 1977). The scale effect refers to the resolution by which aggregate statistics of entities, that is, people, is reported. Relevant to the CCRI is the consideration of what geographic identifiers will be provided with microdata records as well as what geographic framework to provide for researchers to conduct analyses and report findings. While accuracy of geographic location for records can go down to the street level, practical reasons that include the generalization of results and the protection of anonymity require scales that are more generalized.

The aggregation or zonation effect can be particularly challenging. Given the large number of ways that any given space can be partitioned, what is the optimal way and what are the implications of partitioning a space in that way? To what degree is the notion of optimum dependent on the intentions of the researcher? Census geography is the de facto for spatially referencing census aggregate data. Since census geography follows municipal and other administrative boundaries in many cases, analyses need to recognize the limitations imposed by such a framework on the interpretation of findings.

Figure 2, reproduced from Amrhein (1995) demonstrates the effect of the modifiable areal unit problem. In the figure, partitions (a), (b) and (d) demonstrate that a different number of partitions can influence the mean that is calculated. Also, when partition (b) is compared with partition (c) or, partition (d) with partition (e), it can be seen that even given the same number of zones, the orientation of the zones can also produce different results for the mean. The example effectively demonstrates the scale and aggregation effects though not given a context, it may seem somewhat artificial. To bring to mind an example of when this may be occur in a practical research example, consider the case in which a researcher is studying the migration patterns of immigrants. In addition to producing different results for sample point estimates, the zonation effect can altogether obscure significant patterns. Figure 3 shows a space partitioned in two ways with partition (a) composed of zones that are horizontally oriented and partition (b) composed of zones that have a vertical bias. The arrows show the dominant migration pattern.
Notice that since the dominant migration flow is in a vertical orientation, partition (b), unlike partition (a), would not register any movement.

Figure 2 Modifiable Areal Unit Problem. Reproduced from "Searching for the elusive aggregation effect: evidence from statistical simulations" (Amrhein 1995).
In terms of statistical implications, the scale effect was shown to increase the correlation between variables as aggregation increased. The aggregation effect showed that the various combinations of zones that are possible for representing a region can produce point estimates that vary across the gambit of possible values. Certainly some constraints are present, which can reduce an otherwise confounding problem to a smaller subset of possibilities. For example, if zones are constructed from basic units of census geographies, construction of zones will be limited to those which can be formed from the rudimentary units. The point to take away is that any given geographic representation is only one of different possible models, as such, results can vary from one set of zones to the next. This underlies errors of ecological fallacy. Since zones can be defined in different ways and the point estimates of sampled populations can vary between zone definitions, it should be clear that the generalizations extracted from analyses are applicable only to the zones for which point estimates are calculated. In other words,
findings made from a given set of aggregations should not be assumed to hold constant across other possible aggregations and taken to be a universal characteristic of the population (Johnston, Gregory et al. 2000).

The modifiable areal unit problem is therefore an intrinsic problem in the construction and use of microdata. Both effects are embedded in the methodological considerations for the creation of a geographic framework. Decisions made about the scale and zone definition for the geographic framework can have explicit impacts on the outcome of subsequent research. There is a notion of an optimal range when scale considerations are made. The scale effect is bounded by the need to ensure anonymized reporting of results on the one end, and the limits of highly correlated variables on the other. The aggregation effect is in part bounded by the level of geographic precision of the data source, but the possibilities for how areal zones can be defined to partition space are operationally broad. The problem is open to investigation particularly as it applies in the endeavour to construct microdata.

The creation of aggregation zones for microdata requires considering both logistics and politics. The logistics involve guidelines and rules to meet anonymity requirements. The politics involve including the local and regional perceptions and knowledge of local histories. A relevant case study of this problem involves the creation of the Public Use Microdata Areas (PUMAs). PUMAs were created as the reporting unit for census microdata in the United States and they are constructed from aggregations of smaller census geographic entities (see Plate 6 and Plate 7). The U.S. Census Bureau encountered state opposition to the U.S. Census Bureau defined PUMAs, initially released for the 1990 Census. For the 2000 census, the Census Bureau gave each state the opportunity to define their own PUMAs with prescribed guidelines so that local history and socio-economic knowledge could be incorporated to make the zones more meaningful for the states.
Counties are the primary legal divisions of most states. They are delineated cooperatively for statistical purposes by the U.S. Census Bureau. County subdivisions are the primary divisions of counties and statistically equivalent entities for the reporting of decennial census data. They include census county divisions, census subareas, minor civil divisions (including barrios and barrio-pueblos in Puerto Rico), and unorganized territories. (U.S. Census Bureau, 2004)

Census Tracts

Census tracts are small, relatively permanent statistical subdivisions of a county delineated by local participants as part of the U.S. Census Bureau's Participant Statistical Areas Program. The U.S. Census Bureau delineated census tracts in situations where no local participant existed or where local or tribal governments declined to participate. The primary purpose of census tracts is to provide a stable set of geographic units for the presentation of decennial census data. The 2000 Census is the first decennial census for which the entire United States is covered by census tracts. (U.S. Census Bureau).

U.S. Central Cities & Incorporated Places

This diagram shows the relationship between central cities, census subdivisions, and incorporated places. Where Census subdivisions are defined by census county divisions (CCD), census subareas (Alaska), minor civil divisions (MCD), unorganized territories, and incorporated places. CCDs are defined for areas not covered by MCDs and incorporated places.

Block Groups

A census block group (BG) is a cluster of census blocks having the same first digit of their four-digit identifying numbers within a census tract. For example, block group 3 (BG 3) within a census tract includes all blocks numbered from 3000 to 3999. BGs generally contain between 600 and 3,000 people, with an optimum size of 1,500 people. (U.S. Census Bureau)
A public use microdata area (PUMA) is an area with a decennial census population of 100,000 or more people for which the U.S. Census Bureau provides specially selected extracts of raw data from a small sample of long-form census records screened to protect confidentiality. These extracts are referred to as "public use microdata sample (PUMS)" files. Data users can use these files to create their own statistical tabulations and data summaries. For Census 2000, there are state-level and national PUMS files. The U.S. Census Bureau first provided PUMS information in conjunction with the 1960 census data tabulations. (U.S. Census Bureau)

PUMA Delineation Rules

1. All areas in a state must be completely contained by a 1% or 5% PUMA.
2. Each 5% PUMA must always contain 100,000 people.
3. Counties with fewer than 100,000 people cannot be more than one PUMA.
4. PUMA boundaries can only follow the following geographic areas:
   - Counties and statistically equivalent entities.
   - Minor Civil Divisions (MCDs), but only in the New England States.
   - Census tracts, but only within counties that have more than 100,000 people.
   - Places with a population of 100,000 people and more. The use of incorporated place boundaries is only permitted when the PUMAs for these places comprise only the area of these places.
5. Each PUMA must constitute a geographically contiguous area.

PUMA Delineation Recommendations

1. Wherever possible, each PUMA should comprise an area entirely inside or entirely outside MAs.
2. Wherever possible, a PUMA should not cross the boundaries of metropolitan statistical areas (MSAs), primary metropolitan statistical areas (PMSAs), or consolidated metropolitan statistical areas (CMSAs).
3. The Census Bureau recommends the delineation of one or more PUMAs to exclusively contain the entire area of incorporated places that have a population of 100,000 or more.
4. The Census Bureau recommends that the number of 5 percent PUMAs be maximized, and that 5-percent PUMAs should not contain more than 200,000 people, wherever possible.
5. Wherever possible, the 5% PUMA boundaries should conform to the 1990 census 1% PUMA boundaries in order to provide data users with comparability.

Source: U.S. Census Bureau

$ Source: http://www.census.gov/geo/puma/puma_guide.pdf
Justin Ngan, 2004
3.3. The Complicity of Geography

A geographic framework can provide improved contextual understanding for distributions occurring in and statistics calculated for local or regional areas. That same geography however, can also be complicit in creating challenges for the researcher. A geographic framework relying on such ephemeral definitions as political and operational boundaries will make it difficult for the researcher to compare results from census to census. If boundaries change, as they commonly do, analysis is made more difficult because change in the measured statistic will then be partitioned into a component of change resulting from the process of interest and a component of change resulting from shifting boundaries (Ellis, Reibel et al. 1999).

One of the more common situations facing researchers working with geographically partitioned data is the problem of the split census tract. Since census tracts, both in Canada and the United States, are created to maximize socio-economic homogeneity within an ideal population threshold\(^7\). They are prone to redefinition between census periods, especially in rapidly growing areas. Methods of dealing with split tracts between census periods include simple aggregation, which leaves census tracts with a large variance in population sizes, or a combination of aggregation and partitioning of population into split tracts using an underlying assumption such as even distribution or proportional growth (Howenstine 1993). The point here is to note that geography can help by providing context, but that its ephemeral nature especially over the period of some fifty years, needs to be understood and a method developed to enable researchers to conduct comparisons without the need to scrutinize each individual case.

4. Canadian Century Research Infrastructure - Considerations

The CCRI benefits from the experience of its principle investigators who have guided related projects and from the consultation of the knowledgeable creators of the American Integrated Public Use Microdata Series (IPUMS). The decision to bring onboard at an early stage a group to work on the geographic framework will benefit the project through synergies generated from the sharing of cross-disciplinary discussions on how best to address some of the opportunities and challenges that I have introduced in this paper. Already, the team has embarked on the challenging task of recreating the historical census geographies, an effort that requires one to be at least part geographer and part historian. Geographic harmonization of the IPUMS microdata from 1940 to 2000 will be made possible because of the reconstruction of historical census boundaries as part of the National Historical Geographic Information System project. While not a core part of the CCRI mandate, the early inclusion of the geographic team offers the possibility that a

\(^7\)Census tracts in the United States have between 1,500 and 8,000 people with an optimum of 4,000. Census tracts in Canada should have between 2,500 and 8,000 people with a preferred average of 4,000. Note that as of the 2000 Census, the whole of the United States has been partitioned with census tracts. In Canada, census tracts are only defined in Census Metropolitan Areas and Census Agglomerations.
substantive and harmonized geographic framework spanning the entire series will be available for researchers.

I will summarize the ideas introduced in this paper into a series of questions that are more immediately related to the work being undertaken for the CCRI. First, at what spatial resolution should microdata records be sampled? Sampling at small scales, meaning sampling from relatively larger spatial units, might be attractive to those for which the addition of detail is superfluous but, as suggested, that additional detail can be important both in the construction of the sample and in the interpretation of results.

Sampling at large scales, or smaller spatial units, can be incorporated into the design by stratifying the sample by geography. Stratification can increase the precision of sample estimates and overcome the negative effects associated with clustering techniques (Ruggles, Hacker et al. 1995). Given the cluster sampling strategy likely to be employed by the CCRI, a well planned geographic stratification can help reduce standard errors. Stratification by any characteristic can improve the sample derived, but without a priori knowledge of the population distributions, stratifying by geography is not only an effective treatment, but also one of the more viable ones. Geography is embedded in the structure of the manuscript microfilm reels making stratification possible without substantial before hand preparation. From the perspective of interpreting results, a geographic framework with relatively fine resolution, will provide local and regional context that can provide unique insight.

Once the microdata sample has been compiled, the CCRI will face both demands and constraints regarding the level of geographic identifiers attached to the sample records. The benefit of fine scale geographic identifiers in unlocking spatial patterns and processes for researchers must be weighed while respecting the original convenant made between government and people in ensuring anonymity of responses. The decisions made about how to release the data and the level of geographic identifier provided may renew concerns about what understanding the respondents to censuses of the past had about their privacy and how that understanding is interpreted today. The information contained in the census is invaluable, but the value of this ritual may be compromised if respondents of future censuses no longer feel that there is any guarantee of anonymity.

The notion of analysis and reporting leads to the next question. How should the CCRI approach the issue of a temporally harmonized geography for the analysis and reporting of microdata? The need for a consistent geographic framework spanning at least a significant part of the data series would be valuable to the researcher interested in patterns and changes over a period of time. Creating such a consistent framework poses a few issues. Areas as defined by some measure of importance today may not be of similar significance or value in the past. Given the relatively concentrated nature of Canada’s population, it may be difficult to extend backwards in time, a consistent framework of a fine scale that meets requirements for anonymity. In considering these questions, it becomes evident that it will also be necessary to consider the limits to analyses that result from the need to preserve anonymity. Specifically, it will be necessary to examine the limits to statistical measures as a result of aggregating individuals over a large area. In
such cases, to what extent does aggregation increase correlation between reported variables and influence the outcome of findings? The answer to this and similar questions of geography can influence both the development of the CCRI microdata and how that data will be used for research purposes.

5. Conclusion

The Canadian Century Research Infrastructure is an exciting cross-disciplinary project that will yield research undertaken from different social science perspectives to increase the knowledge of Canada’s past and present. The collaborative effort of the various social science disciplines is a great step taken towards the eventual success of the project. In this paper, I have attempted to introduce some of the considerations as seen from a geographic perspective. In doing so, I hope to have brought to mind some questions that might warrant examination now. These questions include considerations of what should be the spatial resolution(s) for the sampling framework, how the geographic framework may be harmonized for the data series, and what the implications of aggregation are to the analyses that will occur given the relatively small population in the early part of the twentieth century. These questions are highlighted for the implications they hold both in the construction phase of the project and in the post-completion phase when the data is used by the research community. It is recognized that the geographic perspective is only one of numerous ways to approach the project. It is one that highlights the relationships between people and space, and that those relationships can be explored to gain an advantage in the creation of and value derived from the impressive undertaking of the CCRI.
<table>
<thead>
<tr>
<th>Project</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IPUMS-USA</strong>&lt;br&gt;Minnesota Population Center, University of Minnesota&lt;br&gt;<a href="http://www.ipums.org/usa/index.html">http://www.ipums.org/usa/index.html</a></td>
<td>• 1850 – 1990 • 25 high precision samples of the American population • microdata access • tabular data with various georeferencing unit codes by county, PUMA, Planning Service Area • descriptions of sampling strategies used: • cluster sampling to concurrently build household level data as individual level sampling is completed • stratified sampling (by geography (pre-1960) and also by ethnicity and household size (1960’s on) • oversampled for target groups (black population, hispanic) • Retrieval includes geographic identifiers (census region and division, state, county, civil division, metropolitan area where applicable, city, PUMA, superPUMA, and political ward)</td>
</tr>
<tr>
<td><strong>International IPUMS</strong>&lt;br&gt;Minnesota Population Center, University of Minnesota&lt;br&gt;<a href="http://www.ipums.umn.edu/international/index.shtml">http://www.ipums.umn.edu/international/index.shtml</a></td>
<td>• Harmonizing census microdata from around the world • China, Colombia, France, Kenya, Mexico, United States, Vietnam, Brazil, Ghana • Begun 1999 and still under</td>
</tr>
<tr>
<td><strong>National Historical GIS</strong>&lt;br&gt;Minnesota Population Center, University of Minnesota&lt;br&gt;<a href="http://www.nhgis.org/">http://www.nhgis.org/</a></td>
<td>• U.S. based • Purpose is to compile and enrich U.S. Census summary data and incorporate into a geographic informations system framework negotiation • Creation of web based access • Geographic area work (census tract boundaries) cross referenced to Minnesota Population Center (MPC) data • Will need to contact MPC to obtain methodology and research related to this endeavor</td>
</tr>
<tr>
<td><strong>Public Use Microdata Area (PUMA)</strong>&lt;br&gt;U.S. Census Bureau&lt;br&gt;<a href="http://www.census.gov/geo/www/cob/pu_metadata.html">http://www.census.gov/geo/www/cob/pu_metadata.html</a></td>
<td>• Areas with census populations of 100,000 or more • First used by U.S. census bureau in 1960 • Super PUMAs contain 400,000 people or more • Aggregate census data 1790-2000 • Tied microdata available for 1% sample sizes • 1% sample size based on microdata size of super PUMAs • Cannot be more than 1 state (or statistically equivalent entity) • Entirely within or entirely outside metropolitan areas • Entirely within or entirely outside central cities of metropolitan areas • Creation methodology</td>
</tr>
<tr>
<td>Great Britain Historical GIS Project University of Portsmouth <a href="http://www.geog.port.ac.uk/gbhgis">www.geog.port.ac.uk/gbhgis</a></td>
<td>• GIS integration • project since 1994 • various statistics available 1841-1939 (not microdata form)</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Canadian Families Project University of Victoria <a href="http://web.uvic.ca/hrd/cfp/">http://web.uvic.ca/hrd/cfp/</a></td>
<td>• Sociological perspectives of family including class, ethnicity and religion of Canada though with limited underlying ability for geographic analysis</td>
</tr>
<tr>
<td>International Microdata Access Group University of Montreal <a href="http://prdh3.demo.umontreal.ca/dillon/IMAG/participants.html">http://prdh3.demo.umontreal.ca/dillon/IMAG/participants.html</a></td>
<td>• Bibliography compiled from microdata studies including: sampling strategies, errors, inherent biases</td>
</tr>
<tr>
<td>Sample of Anonymized Records (SARs) Cathie Marsh Centre for Census and Survey Research <a href="http://www.ccsr.ac.uk/sars/">http://www.ccsr.ac.uk/sars/</a></td>
<td>• a.k.a. microdata for UK • recent data (1991, 2001)</td>
</tr>
<tr>
<td>Source</td>
<td>Summary or Description</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Global Review of 2000 Round of Population and Housing Censuses: Mid-decade assessment and Future Prospects United Nations, Statistics Division <a href="http://unstats.un.org/unsd/demog">http://unstats.un.org/unsd/demog</a></td>
<td>• Summary site for census issues including: o Data collection o New technologies o Census mapping • Census mapping section summarizes issues: o Poorly timed execution of mapping exercises with respect to census o Treatment of mapping as a completely separate exercise from the census o Incomplete mapping o Generation of poor frames for surveys and lack of a solid base for developing the GIS • Not historical however,… • The importance of geography (mapping) in the census process is underscored • While information reported is from a high level perspective of census enumeration, construction, and dissemination, the site holds value in terms of contacts with researchers to understand the basis of their findings, and how those underlying principles can be applied to the historical reconstruction of pums • In particular, note “Deficient Enumeration frame” section of symposium paper at <a href="http://unstats.un.org/unsd/demog/docs/symposium_08.htm#_Toc7409273">http://unstats.un.org/unsd/demog/docs/symposium_08.htm#_Toc7409273</a> identifying need for “[historical comparability], capability to produce data for smaller areas, flexibility to produce data for various systems of administrative and statistical subdivisions”</td>
</tr>
<tr>
<td>Inter-University Consortium for Political and social Research Institute for social research, University of Michigan <a href="http://www.icpsr.umich.edu">www.icpsr.umich.edu</a></td>
<td>• Mission includes: o Acquire and preserve social science data o Provide open and equitable access to these data o Promote effective data use • Electronic cross-referenced bibliographic database of publications • Non-geographic perspective in the provision of data but some implicit value (or perhaps pitfalls): o Sampling design issues addressed in context of studies o Broad spectrum of topics that are researched for understanding potential needs and new possibilities given geographic framework</td>
</tr>
<tr>
<td>National survey of families and households (US) National Survey of Families and Households Center for Demography, University of Wisconsin <a href="http://www.ssc.wisc.edu/nsfh">http://www.ssc.wisc.edu/nsfh</a></td>
<td>• Contemporary collection of recent data with three ‘waves’ of data collection completed • Interview based not historical microdata (search: “public use microdata”; • Reference is made to the availability of geographic reference for data under proviso of confidentiality. From this is construed the possibility of benefit in contacting project principals / analysts to determine the degree of integration of geographic considerations in their sampling strategies • Potential value in examining methodologies for over-sampling by race and social criteria (e.g. family structure, household) and for comparability between samples</td>
</tr>
<tr>
<td>Historical Microdata Around the World <a href="http://www.rhd.uit.no/nhdc/micro.html">http://www.rhd.uit.no/nhdc/micro.html</a></td>
<td>• Links to microdata projects around the world (IPUMS, CFP, aHDS) • Some projects listed are data repositories for open contributions • Many microdata projects do not have geographic reference capability</td>
</tr>
</tbody>
</table>
References


University of Essex (2003). Developing the Collection of Historical and Contemporary Census Data and Materials into a Major Learning and Teaching