Carling, J., ed. (2002). *Nordic demography: Trends and differentials.* Scandinavian Population Studies, Volume 13, Oslo: Unipub/Nordic Demographic Society. 7–30.

The Census in global perspective and the coming microdata revolution

ROBERT MCCAA
Minnesota Population Center,
University of Minnesota
rmccaa@umn.edu

STEVEN RUGGLES
Department of History,
University of Minnesota
ruggles@hist.umn.edu

Abstract On the two hundredth anniversary of the first census of Norway, Denmark, and Iceland it is important to recall the history of the census and of — what to many is a little-known resource for population research — census microdata. The population census became universal only in the last half of the twentieth century. Now, anonymized census microdata is beginning to be recognized as a valuable new source for researchers and policy makers. From a review of practices in the United States and elsewhere, this paper argues that issues of statistical confidentiality and standards for the use of census microdata are rapidly being resolved and that a revolution in usage of these valuable data is already underway.

Keywords: POPULATION CENSUSES, CENSUS MICRODATA, CONFIDENTIALITY.

INTRODUCTION

The bicentenary of the first census of Norway, Denmark and Iceland offers an occasion to look back on the history of the census, and forward to the emergence of census microdata as a valuable tool for social and policy research. Only in the last half of the twentieth century did population censuses became universal. Will census microdata samples become universal in the first half of the twenty-first? In the 1990 round of censuses (1985–1994), of 153 countries with populations of one million or more, 134 conducted enumerations. Ninety-four per cent of the world's population was counted (table 1). Fifty-four countries provided researchers access to census microdata, that is anonymized census samples of individuals and households. Some countries restricted access to a single investigator or research facility, but what is remarkable about the 1990s is not only the globalization of the census, but the growing acceptance of anonymized census microdata as statistical instruments (Dale, Fieldhouse, and Holdsworth, 2000). Once requirements of statistical confidentiality are assured, such samples are suitable for distribution to researchers. The availability of census microdata samples for more

Table 1. The spread of censuses and census microdata availability 1950–2000

Census round (centered on 0)	Number of countries conducting a census	Proportion of World's population enumerated (per cent)	Number of countries offering census microdata samples to researchers
1950s	86	79	2
1960s	117	91	27
1970s	124	71	44
1980s	135	94	54
1990s	134	94	54
2000s	146	97	?

Note: There were 153 'countries' with one million or more inhabitants in 2000 (according to political designations in 2000). For the 2000 round censuses (1995–2004) the number of countries is provisional. *Sources:* Population Reference Bureau (2000); United States Bureau of the Census (2000), United Nations Statistics Division (2001); Kelly Hall *et al.* (2000).

than 50 countries is now driving an effort to harmonize census microdata chronologically and spatially. The IPUMS (Integrated Public Use Microdata Series) international project proposes to integrate census concepts and codes at the microdata level for both contemporary and historical census samples, including those of Nordic countries, and distribute custom-designed databases to *bona fide* researchers via the Internet.

The globalization of the population census was due in great part to the widespread recognition of its public utility and to the development of international standards. The value and practicality of census microdata, on the other hand, are only now being recognized. Standards for this new resource are still emerging. Until the 1990s, technology had been the principal obstacle to the dissemination of census microdata, but now, with the exponentially increasing power of desktop computing and the plunging costs of data-storage, the major problem has become matters of policy rather than technology.

If census microdata are to become widely used, issues of statistical confidentiality must be resolved to the satisfaction of the national statistical agencies and the public as well as researchers. Eurostat sponsored five international conferences on the subject over the past decade. Thanks in part to these efforts and others, the standard practice is now to release microdata samples to researchers. Among the 52 member-states in the International Monetary Fund's General Data Dissemination System, almost three of every four disseminate census microdata samples, in one guise or another (International Monetary Fund 2001). The development of international microdata standards will increase further the availability of census samples, thereby facilitating comparative research, both in time and space. Everywhere that public dissemination policies have been adopted, a quantitative and qualitative revolution in research has resulted. Yet, there has not been a single instance of an allegation of a breach in statistical confidentiality. The Nordic countries, leaders in producing nineteenth century census data, have now begun to make census microdata available to academic researchers for the last decades of the twentieth century.

POPULATION CENSUSES BECOME UNIVERSAL

What is a census? Goyer's authoritative international inventory identifies seven characteristics of the modern population census: legal authority by a national government, definition of the area to be enumerated, complete coverage, individual enumeration, simultaneity of enumeration, periodicity, and publication and dissemination of results (Domschke and Goyer 1986). According to this formal definition, the Nordic countries have better claims than any other area of the world to the distinction of conducting the first modern censuses.

Pre-modern censuses were taken by minor political authorities on an infrequent basis for purposes of taxation, military recruitment, forced labor or religious conformity. Undercounts were common, and often only adult males were enumerated. Results were rarely published or disseminated in any form. Indeed, in the Nordic region, there were earlier censuses taken but the results were not published for military reasons. In the seventeenth century, and then increasingly in the eighteenth, counting the population became an increasing concern in the Western European cultural region. Yet, the places enumerated were often limited to populations numbering in the tens of thousands (Anderson 2001). The Nordic countries were the path breakers with nation-wide census enumerations beginning in Sweden in 1750. The Spanish Crown, in the 1770s and again in the 1790s, attempted to enumerate its American millions, but the returns were never fully compiled or published. In 1790, the United States began an enviable record of decennial censuses, yet it cannot lay claim to primacy because a substantial fraction of the population was not enumerated as individuals until after the abolition of slavery more than seven decades later. Other countries' pretensions to first place are weakened by a later start, the failure to publish results or the absence of periodicity.

From the first decade of the nineteenth century, the Nordic countries were joined in taking censuses by only a handful of other countries, and the total population enumerated was below one hundred million. By the mid-century, the number of places taking a census tripled, and the number of people counted approached two hundred million. In 1872, with the first census of the Indian subcontinent by the British colonial authorities, the population enumerated in that decade approached one billion. By 1900 substantial numbers of European and American nations were well on the way toward conducting periodic censuses, thanks in part to organizational efforts of the International Statistics Institute from its founding in 1885.

In the first half of the twentieth century, the emerging universalization of the census was interrupted by two world wars and a global depression. Only in recent decades has the taking of national censuses become a global reality. Consider the record of 153 countries with inhabitants of one million or more in 2000. As a group they comprised six billion people (compared to fewer than twenty million in the fifty excluded microstates). With the end of the Second World War, the 1950 round of censuses re-established the pattern toward increased census taking with 86 of 153 countries conducting enumerations. As a group they accounted for more than three-quarters of the population of the globe (table 1). The 1960 round surpassed this achievement with 91 per cent of the world's population counted and enumerations held in 117 countries. In the 1980 round, a record was set that still stands with censuses in 135 countries covering 94 per cent of the world's population. The 1990s fell short, due to political instability, economic exigencies, and evolving administrative priorities. While the 2000 round promises to exceed even the record of the 1980s, it should be noted that census anniversaries have passed without an enumeration in seven large countries — Russia, Nigeria, Germany, Ukraine, Poland, Tanzania and Afghanistan — totaling more than 10 per cent of the world's population.

If we assume that the 2000 round goes as planned, enumerations will have occurred every decade over the past half-century in 66 countries, accounting for half the world's population. Missing a single enumeration over the same period were forty-three countries, with 38 per cent of the world's inhabitants in 2000. Thirty-two states missed two or three decennial censuses. Only twelve countries, comprising less than 2 per cent of the globe's population, conducted fewer than three enumerations.

Comparability of census data increased greatly over the same period, thanks to the efforts of international, regional, and national statistical organizations, such as the United Nations Statistics Division, the International Statistics Institute, the Latin American Center for Demography (CELADE), the United States Census Bureau, and national statistical agencies in almost every country. Convergence was due to the growing technical sophistication of census operations and the preservation of institutional memory in publications and the archiving of documentation.

Consider as an example the topics covered in censuses in twenty-three selected Asian countries (3.4 billion people in 2000) over the period 1950–1980. Of twenty-five topics identified by the Statistics Division as general population questions to be investigated, half were included in the 1970 round of censuses of all but one or two of the countries (table 2).

Table 2. Topics in censuses of 23 selected Asian countries representing three billion people, 1950–1980.

Census Topic	**1950**	**1960**	**1970**	**1980**
Countries Enumerated	16	21	20	19
Social				
Sex	**16**	**21**	**20**	**19**
Age	**16**	**21**	**20**	**19**
Marital Status	**15**	**19**	**20**	**18**
Family relationship	**12**	**18**	**20**	**17**
Language	8	9	9	8
Citizenship	10	12	14	12
Ethnicity/race	9	11	10	9
Religion	10	15	14	10
Education-related				
Literacy	**13**	**18**	**19**	14
Years of schooling	**14**	**21**	**20**	**19**
School attendance	8	9	13	14
Educational qualifications	2	5	9	7
Geographical				
Birthplace	10	15	14	10
Residence	11	14	**16**	**15**
Duration of residence	5	7	**15**	10
Prior residence	2	5	**15**	11
Urban-rural	11	**18**	**18**	**17**
Demographic				
Children ever born	6	12	**17**	**15**
Children living	2	7	12	13
Economic				
Activity status	**14**	**17**	**18**	**18**
Occupation	**14**	**20**	**19**	**18**
Industry	**12**	**19**	**19**	**17**
Employment status	11	**19**	**19**	14
Housing	6	14	**15**	**15**
Income	3	3	4	2
Technical				
De facto enumeration	10	10	11	10
Household distinction	6	11	14	14
Post enumeration survey	2	5	6	5
Computerized	4	10	**15**	**15**

Note: **Bold face** indicates topics covered in three quarters or more of the countries surveyed. Countries surveyed: Bangladesh, China, India, Indonesia, Iran, Iraq, Israel, Japan, Jordan, Korea (Republic of), Kuwait, Malaysia, Nepal, Pakistan, Philippines, Singapore, Sri Lanka, Syria, Taiwan, Thailand, Turkey, United Arab Emirates, and Vietnam.
Source: Domschke and Goyer (1986).

Universally adopted were four social variables (sex, age, marital status and relationship to householder), two education variables (literacy and years of schooling), and four economic variables (activity status, occupation, industry, and employment status). Income was rarely asked. More common, although not requested in a majority of countries, were questions on educational qualifications, ethnicity/race, language, and number of living children. Two-thirds or more of the countries collected information on housing, number of children everborn, school attendance, religion and citizenship as well as a variety of migration indicators. While no country followed international recommendations to the letter, many national statistical agencies respected basic census concepts. As a result the scientific quality of enumerations was enhanced and chronological and spatial comparability increased.

LIBERATING CENSUS MICRODATA

> ...official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honor citizens' entitlement to public information.
>
> United Nations Statistical Commission
> Report of the Special Session, New York, E/1994/29

In the 1960s a new statistical data source, the public use census microdata sample, was made available to researchers for the first time. In an effort to meet the needs of scholars who required specialized tabulations, the United States Census Bureau created the first public microdata sample: a 1 in 1000 extract of the raw individual data used to create tabulations for the published census volumes (U.S. Bureau of the Census 1964). The 1960 public use sample revolutionized analysis of the American population and led to an explosion of new census-based research. Not only did it allow researchers to make tabulations tailored to their specific research questions, but it also allowed them to apply new methods to the analysis of census data, especially multivariate techniques.

For the 1970 census of the United States, the 1-in-1000 density of the previous census sample was increased dramatically; the Census Bureau provided six independent public use samples, each of which had a 1-in-100 density. Users who required an exceptionally large number of cases could combine the samples to obtain a six percent density, or about 12 million person records. In addition, the 1970 samples provided a variety of alternate geographic codes, although the Census Bureau still did not identify any places with a population of less than 250,000.

In conjunction with the 1970 public use samples, the Census Bureau released a new version of the 1960 sample. Bureau statisticians enlarged the sample density from 1-in-1000 to 1-in-100, and at the same time reorganized the coding schemes and record layouts to be compatible with the samples from 1970. This compatibility made it relatively easy for investigators to pool data from 1960 and 1970, and thus incorporate change into their analyses.

By the late 1970s, the public use samples for 1960 and 1970 had become one of the essential tools of American social scientists. It was in this climate that Halliman Winsborough and a group of other researchers at the University of Wisconsin developed the idea of creating historical public use samples for earlier census years. They obtained funding from the National Science Foundation and contracted with the Census Bureau to create 1-in-100 samples for the censuses of 1940 and 1950.

The Census Bureau also released public use samples for the 1980 and 1990 censuses. These samples include considerably greater geographic detail and subject content than either the 1960 or 1970 public use samples. Therefore, we now have a continuous series of Census Bureau microdata samples for six census years consisting of anonymized records spanning the period from 1940 through 1990. The Bureau will produce a seventh sample for the 2000 census. Meanwhile, historical samples were constructed from the census enumeration sheets for 1850, 1860, 1870, 1880, 1900, 1910 and 1920. In 2002, work will begin on drawing a 1930 sample. The delay is due to a condifientiality law which prohibits the public release of original census enumeration forms, including names and addresses, until the lapse of 72 years. Only the 1890 census, whose forms were destroyed by an accidental fire more than a century ago, will lack a sample (Ruggles and Menard 1995).

Meanwhile in Latin America in the 1960s, the United Nations Center of Demography (CELADE) began the OMUECE project to create and harmonize microdata samples from the original computerized records for all of Latin America. For the 1960 round of census, samples were drawn for sixteen countries, with densities of one to ten percent. A decade later, nineteen countries participated and average sample densities rose to over five percent. The OMUECE project sought to facilitate comparative demographic analysis by integrating census samples at the level of individuals and households. However, before the 1980 round of enumerations was completed, the project was suspended. Hampered by the fact that, on the one hand, computing resources were still relatively expensive, and on the other, distributing the data was inconvenient, slow, and costly, OMUECE samples were restricted to CELADE personnel, national statistical agencies or visiting researchers. The OMUECE project proved the feasibility of the idea of harmonizing census microdata, but it also demonstrated that widespread usage would only become practical with a massive drop in computing and distribution costs (McCaa and Jaspers-Faijer 2000). The microcomputer revolution of the 1990s solved these problems and more.

Census microdata samples for Asian countries have been collected by researchers at the University of Hawaii's East-West Center and at the Australian National University. For the Republic of Korea, the East-West Center assembled a complete collection of quinquennial census samples dating from the 1960s. Beginning in the 1970s, decennial samples for the Philippines, Indonesia and Thailand were constructed along with those for many states of Micronesia. The capstone of the collection consisted of one percent samples for the Republic of China for the 1982 and 1990 enumerations. Unlike the CELADE collection, little effort was made to harmonize the microdata for specific countries or years, and usage was limited to researchers on site (Minja Kim Cho, private communication, April 29, 2000).

For Africa, beginning with the 1970 round, the Population Studies Center at the University of Pennsylvania amassed a collection of microdata for 33 censuses, (see the African Census Analysis Project website: *www.acap.upenn.edu*). Although in the 1960s, 38 of 59 African countries conducted censuses no microdata are known to survive, other than for the Kenyan census of 1969. For the 1970s, ten microdata samples exist from the 49 censuses conducted. From the 1980s fourteen microdata sets are known to survive (51 censuses), compared with only ten in the 1990s (45 censuses; tallies are from the IPUMS International census microdata inventory, see Kelly Hall, *et al.* 2000). Although these valuable microdata have been little used in the past, the African Census Analysis Project has begun to make the samples more widely available to researchers, particularly those of African or African-American origin.

Also in the 1970s, national statistical organizations, such as Statistics Canada, began to disseminate census microdata samples in growing numbers. For example, in the case of Canada, the 1971 revision to the Statistics Act made possible the public release of non-confidential microdata (Tambay and White 2001). Since the 1970s Statistics Canada, with its series of quinquennial enumerations, has regularly issued census microdata samples. Until 1996, researchers had to request samples individually and distribution was highly restricted. In that year a data liberation initiative was instituted to permit Canadian universities to disseminate microdata samples to researchers and their students. The result was an explosion of research. While before liberation five or ten scholars might acquire microdata samples per year, afterward, a single sample at a major university might be accessed hundreds of times per month. The profusion of suppliers means that usage statistics are now impossible to compile, where before the agency recorded every user by name. Given the widespread use of census microdata in the university classroom, Canadian scholars are educating a younger generation of citizens about the utility of the census and democratizing access to census data (Lisa Dillon, private communication, April 21, 2001).

In the United Kingdom, public use census samples called SARs (Samples of Anonymized Records) were first constructed for the 1991 enumeration, with a sample density of 2 per cent for individual records, and 0.5 per cent for households. Administrative units with fewer than 120,000 inhabitants were not identified. Notwithstanding the small density of the samples and the absence of geographical detail there was a great out-pouring of research based on the SARs. Hundreds of studies were published within six years of their initial release. In anticipation of the 2001 enumeration, disclosure risks were re-assessed, now taking into account error and coding variability as well as differences in timing and coding schemes between datasets. With privileged access to a benchmark housing survey, an attempt was made to actually match records against the SARs in order to assess empirical, as opposed to theoretical, risks (Dale and Elliot 2001). The authors reasoned that theoretical studies, which accounted for all prior assessments, exaggerated the risks of identifying an individual because they neglected to take into account error, differences in timing of sources and incompatibilities of coding schemes. From this time-consuming, exhaustive test, Dale and Elliot conclude:

> For a user of an outside database, attempting this sort of match with no opportunity for verification would prove fruitless. In the first place, the small degree of expected overlap would be a considerable deterrent to an intruder. However, if a match between the two files was attempted the large number of apparent matches would be highly confusing as an intruder would have no way of checking correct identification.

For the 2001 SARs, Dale and Elliot propose a halving of the geographical threshold and a fifty percent increase in sample density. In the case of U.K. census microdata, risks are substantially reduced because all data are entirely categorical. The authors note that confidentiality risks increase greatly where finely detailed occupational or geographic codes or interval level variables, such as income, are made available.

To sum up, everywhere census samples have been made public, such as the data liberation initiatives in the United States, Canada and the United Kingdom, the result has been an explosive growth in research based on census microdata, at little risk to the public. With respect to the experience of the United Kingdom Dale and Elliot (2001) conclude their analysis of the statistical confidentiality issue as follows:

> There has been no known attempt at identification with the 1991 SARs — nor in any other countries that release samples of microdata. Our research on the scenarios under which an attempt might occur suggests that there would be no commercial advantage in attempting to make direct matches with an external database and that the main danger comes from maverick attempts to discredit either the census operation, ONS [Office of National Statistics] or the government (Elliot and Dale 1999). Attempts to discredit these bodies would be much simpler using more readily accessible data — for example, by infiltrating the census-taking process.

HISTORY OF STATISTICAL CONFIDENTIALITY IN THE UNITED STATES.

Concern for the confidentiality of U.S. statistical data began in 1850, when the Secretary of the Interior declared that henceforth census returns were to be 'exclusively for the use of the government, and not to be used in any way to the gratification of curiosity, the exposure of any man's business or pursuits, or for the private emolument of the marshals or assistants' (Wright and Hunt 1900). Over the course of the next century, confidentiality procedures became increasingly rigorous. By 1880, enumerators were required to swear an oath not to disclose any information to anyone except their supervisors, and in 1910 the Bureau was prohibited from publishing tables in which the identity of a particular business establishment might be deduced by competitors. But the right to privacy was still not absolute; the Director of the Census had the authority to release data on individuals for worthy purposes. As late as 1921, for example, the Director allowed a private literacy campaign to use the census to identify illiterates (Bohme and Pemberton 1991, General Accounting Office 1998).

In the 1920s, the United States Census Bureau began to deny all access to data on individuals, even when the request came from another government agency. In 1930, the Bureau turned down a request from the Women's Bureau for the names and addresses of employed women, and in 1942 it denied a request from the War Department for the addresses of Japanese-Americans, although it did prepare specialized tabulations to aid in the internment of Japanese-Americans (Bohme and Pemberton 1991, Seltzer and Anderson 2000). This expansive interpretation of the right to confidentiality was codified by Title 13 in 1954, which prohibits 'any publication whereby the data furnished by any particular establishment or individual under this title can be identified' (Title 13 United States Code Section 9).

Ensuring confidentiality was comparatively straightforward as long as government data consisted of printed tabulations of the number of persons who had a particular characteristic or combination of characteristics. The advent of electronic computers, both within the Census Bureau and on university campuses, allowed a groundbreaking shift in this paradigm for the Census of 1960: the release of microdata samples.

The release of individual-level information was not seen as a violation of Title 13 because the Bureau did not reveal the identity of individuals. To preserve confidentiality, the Census Bureau removed names and addresses. The Bureau also suppressed other information that might be used to identify particular individuals. For example, the Census Bureau stripped off all geographic detail below the state level. Income was top-coded to prevent the identification of the very rich. Moreover, the fact that only 1 of every 1000 persons was included in the sample was thought to provide a measure of confidentiality protection.

These provisions ensured statistical confidentiality, but they also imposed severe limitations on the usefulness of the data. The geographic restrictions meant, for example, that it was impossible to identify the population of New York City. The small sample size posed additional problems. The 1 in 1000 sample density yielded about 180,000 person records. Given the modest capacity of computers in 1964, this was a lot of cases, but as researchers began to use the sample for detailed analysis of small population subgroups, its limitations became apparent. To address these problems, in 1970 and again in 1980 the Bureau greatly expanded both geographic detail and sample density, and the census microdata samples became the single most important data source in American social science (Ruggles 2000).

The microdata revolution was not limited to the census. The Bureau created public microdata samples from the Current Population Survey and other data products. Recognizing the power of microdata, government agencies in both the United States and, as we have seen above, other countries followed suit and began to release anonymized individual-level data files to researchers. Between the early 1970s and the early 1990s, the number of such files exploded, and the detail they provided on geographic and personal characteristics expanded dramatically.

The confidentiality protections in microdata released by the Census Bureau and other statistical agencies have been an unqualified success. After discussing the issue with dozens of specialists in the field, Ruggles was unable to confirm a single case of disclosure of the identity of an individual in a public-use research dataset. Thus, despite the free availability of U.S. census microdata for 37 years, we have no evidence that anyone's confidential information has been revealed.

In theory, the identity of an individual in a microdata file might be revealed by matching his or her characteristics to a public or private source that includes names. For example, there might be some sort of private database containing name, age, sex and marital status for a subset of individuals in a locality. By searching an anonymized microdata file for persons with the same combination of characteristics, it might be possible to guess the identity of a respondent. This would result in what analysts have termed *re-identification disclosure* or *inferential disclosure*.

In practice, such disclosure of confidential information is highly improbable, as Dale and Elliot have shown with the 1991 SARs for the United Kingdom. These microdata are samples, and none of them includes information on more than a tiny minority of the population. For this reason alone, any attempt to identify the characteristics of a particular individual, in say a five percent sample, would necessarily fail at least nineteen times out of twenty. Even in the event that one located a unique exact match for a target individual, one could never be certain that the case actually represented that individual; there is always the possibility that there exists another exact match not included in the sample.[1] Moreover, statistical microdata are not designed to support the investigation of particular individuals. Some datasets, including the U.S. census microdata, introduce deliberate alterations of individual characteristics to enhance confidentiality protection. In addition, as Dale and Elliot have shown, statistical microdata are subject to noise resulting from respondent and data processing errors and ambiguities. In sum, identification is impossible for the vast majority of persons, positive identification is always impossible, and statistical data are an inferior source for identifying the characteristics of particular individuals.

A recently published, worst case scenario of potential adverse consequences of respondent disclosure in the United States offers the following list of possible dangers (Mackie, in press):

> Disclosure of personal information may result in an individual being arrested for a crime, denied eligibility for welfare or Medicaid, charged with tax evasion, losing a job or an election, failing to qualify for a mortgage, or having trouble getting into college.

None of these injuries has ever occurred. It is difficult to imagine a scenario under which any of these consequences could result from a breach of the privacy protections in a statistical dataset. For example, most people who lie on their tax returns or their welfare applications would probably also lie to statistical interviewers. If state and federal agencies are interested in detecting fraud, there are far better ways to do it than by employing hackers to try to crack the security measures built into census microdata samples.

Use of statistical microdata by the private sector to breach confidentiality protections is equally implausible. In the United States, privacy is indeed under assault. There are now on the Internet hundreds of web sites promising full investigative reports on any individual — including credit ratings, property records, marital history and other information — for fees ranging from $35 to $150. Given this wealth of information readily available from private sources, it would be foolhardy to turn to statistical microdata to attempt to uncover imprecise and outdated information about a particular individual.

For all these reasons, the unblemished record of privacy protection in statistical data released by the U.S. government should come as no surprise. It makes sense that the methods currently in place for preventing disclosure of information in statistical datasets provide effective data security. Nevertheless, there is growing concern about confidentiality issues within government. Amid a chorus of concern about privacy, government agencies are reducing the detail available in microdata products and developing restricted-access dissemination procedures for many new microdata products.

Why is there heightened concern about confidentiality now? According to the U.S. Census Bureau, faster computers and more sophisticated search software have increased the potential for uncovering identities of individuals in microdata. Moreover, all data-gathering organizations are rightly concerned about declining response rates and about increasing public concerns about privacy.

Most discussions of confidentiality, describe two alternative approaches. First, there are those who are developing methods of restricting access to data. This includes the restricted data enclaves such as the United States

Census Bureau Research Data Centers, restricted use licenses or agreements, and web-based analysis systems that incorporate automatic suppression of small cells. The second approach is to modify the data to minimize the risk of disclosure. These modifications can be quite simple — such as the suppression of geographic detail and top-coding of long-tailed variables — or more complex, including swapping, microaggregation, and other forms of data perturbation.

The problem with data enclaves such as the U.S. Census Bureau Research Data Centers (RDCs) is that they impose heavy costs on social science and policy research. It is not easy to use a Census Bureau Research Data Center. Because of the cost barriers and inconvenience, the RDCs have attracted few researchers. Only well-funded investigators doing work deemed valuable by the Bureau are eligible to use the centers. The user registration logs for the IPUMS data extraction system (*www.ipums.org*) suggest that a majority of microdata users are graduate students, who would for practical purposes be excluded from using an RDC. Even if the funding problem could somehow be overcome, the number of seats in the centers would have to be multiplied a thousand fold to accommodate the current number of users of public microdata. The RDCs were never intended as a substitute for public use microdata and they cannot fulfill that role.

The alternative to restricted access is data modification. The most straightforward data modification is the reduction of detail, but researchers have expressed alarm at this alternative. In May 2000, Ruggles was asked by the Census Bureau to report on the potential impact on users of reducing the detail offered by the 2000 Public Use Microdata Sample (PUMS) of the United States census, which was then being contemplated as a means of reducing the risk of disclosure. Accordingly, approximately 1,300 current users of the IPUMS-USA data were emailed and requested that they fill out a web-based survey on the issue. Within seven days, 1,006 researchers had completed the survey. The reaction was remarkably uniform: data users overwhelmingly expressed a preference for maximum detail, and described hundreds of research projects that would have to be abandoned if the Bureau reduced detail significantly. Many users were outraged by the suggestion that subject detail might be reduced; one wrote, for example, 'As far as I am concerned, elimination of the detail of age, race, ancestry, income, occupation, and geography would essentially eliminate the value of data from the long form. This is a shameful, cowardly, and ludicrous proposal. I hope it will disappear promptly and not be raised again' (Ruggles 2000).

The risk to privacy posed by publicly accessible microdata must be weighed against the social cost of restricting access to information. That cost is high; if the flow of public use microdata is reduced, we can be certain that use of these data to understand social change and plan for the future will decline proportionately. The risk to privacy, however, is very low; indeed, the safety record for public-use government microdata is apparently perfect.

There is cause for worry, however, if there is a public perception that government data are not adequately safeguarded. To the extent that the public believes their responses are not truly confidential, the cooperation of respondents is likely to decline. But this is a public relations problem and calls for a public relations solution, not a technical solution. If all microdata were to be withdrawn from the academic and policy communities, it is highly unlikely that this would restore confidence because public concerns about privacy have little to do with public-use microdata. Surveying several dozen of the countless web sites that complain about the intrusion of the census into privacy rights, Ruggles found that many people are worried about the potential for the *government* to misuse the information, but few discuss the potential for *public* disclosure.

Perceptions are important and data users have as much interest in ensuring a high census response rate as do data producers. But handicapping social scientists and planners by withholding data is unlikely to allay public distrust of government data collection efforts. If reason prevails, the pressure on agencies to withdraw public-use data will recede without doing too much damage to social science infrastructure. We can then look to the new opportunities that have been created by the clamor over confidentiality. The data enclaves, licensing agreements and statistical perturbation methods have the potential to open a wide range of new administrative data to social science and policy applications. In the end, the renewed concern about the risk of disclosure actually has the potential to enhance access to data rather than restrict it.

Emergence of International Norms of Statistical Confidentiality

> 'statistical confidentiality` shall mean the protection of data related to single statistical units which are obtained directly for statistical purposes or indirectly from administrative or other sources against any breach of the right to confidentiality. It implies the prevention of non-statistical utilization of the data obtained and unlawful disclosure
>
> Council Regulation (EC) No 322/97 of 17 February 1997

In the past decade international norms of statistical confidentiality have emerged, as reflected in the Council Regulation No. 322/97 of the European Community. This regulation was the product of regular biennial meetings on the subject, beginning in 1992 at Dublin. These were followed by work sessions in Luxembourg (1994), Bled (1996), Thessaloniki (1999), and Skopje (2001). Here experts considered efforts to strike a balance between safeguarding the data and facilitating use (Holvast 1999). The most recent meeting, held in Skopje in March 2001, discussed the changing role of National Statistical Institutes, from collectors and disseminators of macrodata to distributors of a wide array of microdata (Eurostat Secretariat 2001). This reassessment was underway at the 1999 meeting. There, Thorogood argued for the broader dissemination of microdata because, on the one hand, of their high value, collected at public expense often with a burden on respondent, and on the other, of the public interest in exploiting the data for the public good. He identified seven practices at the European level for safeguarding statistical confidentiality (Thorogood 1999):

- Small sample size
- Limited geographical detail
- Top and bottom coding of unique categories
- Signed non-disclosure agreements
- Prohibition of redistribution of datasets to third parties
- Prohibition of attempting to identify individuals or the making any claim to that effect
- Requiring users to provide copies of publications

The issue of statistical confidentiality has taken on a global dimension. Among the 52 member-states in the International Monetary Fund's General Data Dissemination System, almost three of every four disseminate census microdata samples, in one guise or another. An examination of the language of the provisions of confidentiality reveals a surprising uniformity of concepts, language, and norms. A synthesis of confidentiality provisions of the 52 member-states of the International Monetary Fund's General Data Dissemination System is available online.[2] With the emergence of a consensus at the level of policy and law, semi-automatic methods for assessing the risks to statistical confidentiality in microdata, such as the ARGUS system developed by Statistics Netherlands, may prove useful (Hundepool, *et al.* 1998.

HARMONIZING AND DISSEMINATING CENSUS MICRODATA SAMPLES

As microdata samples become available, scholars will want to use them. One way of increasing use is by standardizing documentation, codes and concepts for census samples to facilitate analysis of change over time and across space.

Harmonizing census data is not a new idea. Michael Drake discovered that the idea was first proposed in 1872 at the International Statistics Congress held in St. Petersburg (Drake 1997),[3] not much progress was made until the last half of the twentieth century. One of the signal achievements of the United Nations Statistics Division has been in the international harmonization of census concepts from the enumeration form to the publication of final tables (United Nations Statistics Division 1947–, 1998). While incomplete, the effort has enjoyed widespread support by statistical agencies around the globe.

In 1992, a project was begun at the University of Minnesota to integrate sixty-five million microdata records for a single country, the United States. Conceived by Steven Ruggles, the National Science Foundation and the National Institutes of Health of the United States funded a series of projects to develop and integrate decennial census microdata of the United States, dating from 1850 to 1990. By 1993, the first version of the Integrated Public Use Microdata Series (IPUMS) database was released on tape and by 1995 via the Internet (Ruggles *et al.* 1995; Ruggles *et al.* 2000). Thanks to the expansion of the Internet, the data distribution problem was easily solved by means of a web site driven data dissemination engine (*www.ipums.org*). The IPUMS database, distributed free of charge via the Internet, quickly established itself as one of the three most frequently cited data sources in population research about the United States.

In 1998, Ruggles was persuaded to extend the project internationally. Colombia was selected as a test-site, thanks, first, to an early, enthusiastic response by the Colombian national statistical authority, Departamento Administrativo Nacional de Estadística (DANE), and, second, to the ready availability of an uninterrupted collection of decennial census microdata going back to 1964. In late 1999 funding was obtained from the National Institutes of Health. Dubbed IPUMS-Colombia (at the same time that 'IPUMS' was re-baptized 'IPUMS-USA'), the project got underway in Bogota in January 2000. The IPUMS-Colombia project differs from the USA effort by, first, adopting international standards, and second, by calling upon a team of demographers and statisticians, all but one of whom are Colombians, to design a harmonious national integration scheme. To the extent feasible, codes are being made entirely compatible with emerging international standards such as the United Nations, *Principles and Recommendations* (1998) and the United Nations Economic Commission for Europe and Statistical Office of the European Communities' *Recommendations for the 2000 Censuses* (1998).

Meanwhile, with major funding secured from the National Science Foundation, a global effort, IPUMS-International, was inaugurated. With the cooperation of national teams of investigators, the IPUMS-International consortium proposes to integrate census microdata for more than a dozen countries, with at least two from each continent. Historical census microdata for Canada, Norway, Great Britain, and Argentina will be included in the database as well as those for the United States. Contemporary microdata for Colombia and the United States will be integrated along with those for France, Brazil, Mexico, Vietnam, China, Kenya, Great Britain, Hungary, Spain, and a growing list of other countries (table 3).

Based on the IPUMS-Colombia prototype, country teams of experienced census data-users are recruited to advise on how to harmonize the national census concepts using international norms.

Table 3. The IPUMS International Consortium (January, 2002).

Country	Censuses	Sample densities (per cent)
Argentina	1869, 1895	5–7
Brazil	1960, 1970, 1980, 1991, 2001	5
Canada	1871, 1881, 1901	1.7–100
China	1982, 1990*, 2000*	0.1–1
Colombia	1964, 1973, 1985, 1993, 2003	1–10
Finland	1950*, 1960*, 1970*, 1975*, 1980*, 1985*, 1990*, 1995*, 2000*	5–10
France	1962, 1968, 1975, 1982, 1990	5
Ghana	1984*, 2000*	1–10
Hungary	1980, 1990, 2001	5
Kenya	1969*, 1979*, 1989, 1999	5
Mexico	1960, 1970, 1990, 2000	1–5
Norway	1801, 1865, 1875, 1900, 1960*, 1970*, 1980*, 1990*, 2001*	2–100
Spain	1981, 1991, 2001	5
United Kingdom	1851, 1881, 1961*, 1971*, 1981*, 1991, 2001	1–100
United States	1850, 1860, 1870, 1880, 1900, 1910, 1920, 1940, 1950, 1960, 1970, 1980, 1990, 2000	1–100
Vietnam	1989, 1999	3–5

Asterisks (*) indicate censuses under consideration

The IPUMS International project has four goals:

1. Inventory machine readable census microdata
2. Preserve census microdatasets identified as at-risk
3. Create an integrated international census database with a harmonized system of concepts, variables and codes, incorporating both historical and contemporary microdata samples of individuals, households and dwellings
4. Disseminate integrated microdata samples via the internet, using techniques similar to the ipums-usa web-based system (www.ipums.org), but, unlike the ipums-usa, restricting access to bona fide researchers who have signed a non-disclosure agreement.

The first task of the IPUMS International project is to inventory census microdata currently known to exist (see Kelly Hall *et al.* 2000 for an up-to-date listing). The second is to preserve those datasets identified as at risk. Of 235 extant sets currently inventoried, almost one hundred are being preserved under the direct auspices of the initiative. We suspect that another fifty still survive, but these remain unverified until physical existence is confirmed by means of a count of the actual number of machine-readable records.

Complete and comprehensive metadata are essential to the success of the International project. For every country where census microdata currently exist, whether they may be integrated into the international database or not, we seek to preserve four types of documentation for each dataset: codebooks, original enumeration schedules, enumerator instruction booklets, and data processing instructions. For the most recent census microdata these materials may be published, indeed in a single volume. For many countries, some of these materials may be elusive for earlier censuses, existing only in archival form, often as typescripts with, at times, only a single surviving copy.

The third task of the project is to make census microdata for selected countries usable (table 4). Although large machine-readable census microdata exist for many countries, access to these data is restricted in virtually every case. Countries that join the IPUMS International integration consortium agree — given appropriate privacy and confidentiality safeguards — to permit access by accredited researchers to samples of their census microdata. The goal of the project is not simply to make international microdata available; it will also make them usable. Even in the few cases where microdata are already available to researchers, comparison across countries or time periods is challenging owing to inconsistencies between datasets and inadequate documentation of comparability problems (Domschke and Goyer 1986). Because of this, comparative international research based on pooled microdata is rarely attempted. The project promises to reduce the barriers to international research by preserving datasets and making them freely available to qualified researchers, converting them into a uniform format, providing comprehensive documentation, and developing new web-based tools for disseminating the microdata and documentation.

The integration project is composed of four interrelated elements. The first is planning and design. The international dimension of the database poses new design challenges, since it must accommodate variations in census design and cultural concepts. The starting point for developing an integrated design must be the standard classification schemes in the field of international population censuses, including, but not limited to, the following:

o United Nations Statistics Division (1998) *Principles and Recommendations for Population and Housing Censuses.*

o UNESCO (1997) *The International Standard Classification of Education (ISCED 1997).*

o International Labor Office (1990) *International Standard Classification of Occupations (ISCO-88).*

o United Nations Statistics Division (1990) *International Standard Industrial Classification of All Economic Activities (ISIC-88).*

The basic design goals remain the same as in the IPUMS-USA: the international system should simplify use of the data while losing no meaningful information except where necessary to preserve statistical confidentiality.

The second element, microdata conversion, falls into two categories. For some countries, such as China, France, Kenya, Vietnam, Mexico and Brazil, the project will incorporate already-existing samples. For other countries, no accessible census files presently exist (e.g., Spain, and for microdata prior to the 1990s, Colombia, Great Britain, and Hungary). In these instances, new samples will be drawn from surviving census tapes using techniques to ensure that respondent confidentiality is preserved. These data files are often not publicly documented and require extensive assistance from the statistical offices and experts of each country to assure their correct interpretation.

The third element, the development of metadata, is central to the project and poses even greater challenges than the microdata. The documentation is not confined to codebooks, census questionnaires and enumerator instructions. As with the IPUMS-USA, a wide variety of ancillary information will be provided to aid in the interpretation of the data, including full detail on sample designs and sampling errors, procedural histories of each dataset, full documentation of error correction, anonymization procedures and other post-enumeration processing, and analyses of data quality.

The final element of the project is the creation of an integrated data access system to distribute both the data and the documentation on the Internet. With the IPUMS international access system users will extract customized subsets of both data *and* documentation tailored to their particular research questions (unlike the IPUMS-USA system, where the entire documentation system is provided to the user, regardless of the data requested). The IPUMS International system will consist of a set of tools for navigating the mass of documentation, defining datasets, and constructing customized variables. Given the large number of variables and samples, the documentation will be so unwieldy as to be virtually unusable in printed form. Accordingly, the project is developing software that will construct electronic documentation customized for the needs of each user.

Variable design often influences the analytical strategies adopted by researchers, and therefore plans must be developed with care. There are two competing goals. On one hand, we must keep the variables simple and easy to use for comparisons across time and space. This requires that the lowest common denominator of detail be provided that is fully comparable, with underlying complexities transparent to the user. On the other hand, all meaningful detail in each sample must be retained to the extent compatible with statistical confidentiality, even when it is unique to a single dataset.

The project is employing several strategies to achieve these competing goals. In some cases, the original variables are compatible and their recoding into a common classification is straightforward. The documentation will note any subtle distinctions a user should be aware of when making comparisons. For most variables, however, it is impossible to construct a single uniform classification without losing information. Some samples provide far more detail than others, so the lowest common denominator of all samples inevitably loses important information. In these cases, we are constructing composite coding schemes. The first one or two digits of the code provide information available across all samples. The next one or two digits provide additional information available in a broad subset of samples. Finally, trailing digits provide detail only rarely available. The data access system will guide researchers to use only the level of detail appropriate for the particular cross-national or cross-temporal comparisons they are making.

In some cases, incompatibilities across samples are so great that the composite coding scheme is significantly more cumbersome than the original variable coding design. In these cases, we are developing alternate versions of the variables suitable for particular comparisons across time and space. The data access system will recommend the most appropriate version of each variable to researchers based on user profile and the particular combination of datasets they are using. We anticipate that this approach will be needed more often in the international context than it was in the construction of the original IPUMS. Where feasible, we base our coding designs on international coding systems. For geographic variables, we generally conform to the standard of the country.

Most data transformations are simple recodes of one value into another. As in the case of the original IPUMS, we are developing data transformation matrices, which provide information on the location of the original variable in each sample, each original data value, and each new standardized data value. These matrices are maintained in a standard relational database. The actual recoding operations, however, will be carried out with a C program operating as a sequential batch process, since that is the most efficient approach with respect to both storage and speed. In many instances, it is necessary to use information from more than one variable in the original to construct a new compatible variable. For example, one might need information on both province and sub-district to identify a metropolitan area. Data transformation matrices can sometimes handle such complex transformations, but in other cases we must resort to customized programming solutions.

One of the signal contributions of the IPUMS to the original U.S. census files was the creation of family interrelationship variables in all years. Similar variables are being constructed for the international database. A system of logical rules identifies the record number within each household of every individual's mother, father, or spouse, if they were present in the household. These 'pointer' variables allow users to attach the characteristics of these kin or to construct measures of fertility and family composition. For example, use of the spouse pointer variable makes it easy for users to identify spouse's income for each married person in the census. Because of variations across countries in the information available for identifying family interrelationships and in the cultural meaning of marriage (e.g., the high frequency of consensual unions in Latin America and of cohabitation in Scandinavia), we plan to revise the logic of the family interrelationship variables for the international database.

All accessible census microdata files are designed to protect the confidentiality of individuals. Countries have different standards, but in all cases names and detailed geographic information are suppressed and top-codes are imposed on variables such as income that might identify specific persons. Some countries take additional steps, such as 'blurring' a small percentage of geographic information, 'swapping' a small, undisclosed fraction of records (i.e., swapping geographical identifiers), or randomizing the sequence of cases so that detailed geography cannot be inferred from file position.

Many datasets will already have been subjected to confidentiality procedures by the national agency that created the files, and in these cases no additional steps are needed. In a few cases, however, as with Colombia, we are required to work with the original 100 percent machine-readable census returns, from which a nationally representative sample of specified density must be drawn. In such cases, the project is working with each country's statistical office to ensure full confidentiality of all files before they are made available to researchers. Automated methods of evaluating statistical confidentiality, such as the μ-Argus system developed by Statistics Netherlands, will be adopted where possible, with the goal of maximizing available detail while maintaining the highest standards of statistical confidentiality, and the broadest possible policy of dissemination (Hundepool *et al.* 1998).

IPUMS-USA differs from the IPUMS International project in one important respect: dissemination policy. The former is public while the latter is limited to bona fide researchers who sign a non-disclosure agreement. In the case of the United States, the samples released by the United States Census Bureau are public and may be distributed to anyone without regard to issues of statistical confidentiality. The international project distributes integrated microdata of individuals and households only by agreement of the corresponding national statistical offices and under the strictest of confidence. Before data may be distributed to an individual researcher, an electronic license agreement must be signed and approved. To gain access to the data, researchers must agree to the following:

1. Recognize the copyright of the corresponding national statistical agency
2. Use the microdata for the exclusive purposes of teaching, academic research and publishing, and not for any other purposes without the explicit written approval, in advance, of the corresponding national statistical authorities. Researchers must explicitly agree to not use microdata

acquired for the pursuit of any commercial or income-generating venture either privately, or otherwise. It should be noted that the corresponding national statistical authorities may at their discretion approve use for commercial purposes, but not the IPUMS International project.

3. Maintain the absolute confidentiality of persons and households. Any attempt to ascertain the identity of persons or households from the microdata is strictly prohibited. Alleging that a person or household has been identified is also prohibited.
4. Implement security measures to prevent unauthorized access to census microdata. Penalties for violating the agreement include revocation of the license, recall of all microdata acquired, filing of a motion of censure to the appropriate professional organizations, and civil prosecution under the relevant national or international statutes.

Conclusion

Now that the population census has become a global phenomenon, and the construction of anonymized microdata data samples an increasingly widespread practice, harmonization of census microdata is an obvious next step to enhancing use. With the emergence of global standards of statistical confidentiality and the massive power of ordinary desktop computers, the only remaining obstacle is the integration of anonymized census microdata samples. If the experiences of Canada, the United Kingdom and the United States are reliable guides, an explosion in scholarly research is likely to ensue.

Acknowledgements

Research for this paper was funded in part by the National Science Foundation of the United States, grant SBR-9908380 'Integrated Public Use Microdata Series International'.

Notes

[1] Some analysts have argued that individuals can be positively identified if it can be determined from a 100 per cent aggregate summary census file that there is only one individual with a particular combination of characteristics residing in a particular geographic area. However, because small cells in the summary files have also been subjected to confidentiality protection through swapping techniques, they cannot be used to prove that any particular set of characteristics is unique. On the difficulties of matching individuals in a microdata file to a reference file, see Blien, Wirth and Muller (1992) and Dale and Elliot (2001).

[2] See *www.hist.umn. edu/~rmccaa/nordic_appendix.htm.*

[3] Drake cites: Joseph Körösi (1881) *Projet d'un Recensement du Monde.* Etude de Statistique, Paris: 42–43.

References

Anderson, M. (2001) 'Censuses: History and methods.' *International encyclopedia of the social and behavioral sciences.* Oxford: Pergamon.

Blien, U., Wirth, H. and Muller, M. (1992) 'Disclosure risk for microdata stemming from official statistics.' *Statistica Neerlandica,* 46(1): 69–82.

Bohme, F.G. and Pemberton, D.M. (1991) 'Privacy and confidentiality in the U.S. censuses — A history.' Paper presented at: Annual meeting of the American Statistical Association, Atlanta, August.

Dale, A., Fieldhouse, E. and Holdsworth, C. (2000) *Analyzing census microdata.* London: Arnold.

Dale, A. and Elliot, M. (2001) 'Proposals for 2001 SARS: An assessment of disclosure risk.' *Journal of the Royal Statistical Society, Series A,* 164, part 3: 427–447.

Domschke, E. and Goyer, D.S. (1986) *The Handbook of National Population Censuses: Africa and Asia.* Westport CN: Greenwood Press.

Drake, M. (1997) 'Population: patterns and processes.' In: Pugh, M. (ed.) *A companion to modern European history 1871–1945.* Oxford: Blackwell. 3–24.

Elliot, M. J. and Dale, A. (1999) 'Scenarios of Attack: The data intruder's perspective on statistical disclosure risk.' *Netherlands Official Statistics*, Vol. 14:6–10.

Eurostat Secretariat (2001) *Report of the March 2001 work session on statistical data confidentiality.* Joint ECE/Eurostat work session on statistical data confidentiality, Skopje, March.

General Accounting Office (1998) *Decennial Census: Overview of Historical Census Issues.* (Chapter Report, 05/01/98, GAO/GGD-98-103) Washington DC.

Holvast, J. (1999) 'Statistical confidentiality at the European level.' Paper presented at: Joint ECE/Eurostat work session on statistical data confidentiality, Thessaloniki, March.

Hundepool, A., Willenborg, L., Wessels, A., van Gemerden, L., Tiourine, S. and Hurkens, C. (1998) *μ-Argus user's manual.* Voorburg: Statistics Netherland.

International Labor Office. (1990) *International Standard Classification of Occupations (ISCO-88).* Geneva.

International Monetary Fund. (2001) *General Data Dissemination System Bulletin Board.* (Available online at: *dsbb.imf.org/category/popctys.htm.*)

Kelly Hall, P., McCaa, R. and Thorvaldsen, G., eds (2000) *Handbook of international historical microdata for population research.* Minnesota Population Center: Minneapolis. (Updated microdata inventory available at *http://www.ipums.org/international/iiinventory2.html.*)

Mackie, Christopher. (In press) 'Improving confidentiality of and access to research microdata: Summary of a workshop.' *Of Significance — Journal of the Association of Public Data Users.*

McCaa, R. and Jaspers-Faijer, D.J. (2000) 'The standardized census sample operation (OMUECE), 1959–1982 [1992]: A Project of the Latin American Demographic Center (CELADE).' In: Kelly-Hall, P., McCaa, R. and Thorvaldsen, G. (eds) *Handbook of international historical microdata for population research.* Minnesota Population Center: Minneapolis: 287–301.

Population Reference Bureau. (2000) *World population reference sheet.* (Available online at: *www.prb.org/pubs/wpds2000/wpds2000_Population2000-Population Projected.html.*)

Ruggles, S. (2000) 'The public use microdata samples of the U.S. census: research applications and privacy issues.' A report of the Task Force on Census 2000, Minnesota Population Center and Inter-University Consortium for Political and Social Research Census 2000 Advisory Committee. (Available at: *www.IPUMS.org/~census2000.*)

Ruggles, S., Fitch, C.A., Kelly Hall, P. and Sobek, M. (2000) ‘IPUMS-USA: Integrated Public Use Microdata Series for the United States.’ In: Kelly-Hall, P., McCaa, R. and Thorvaldsen, G. (eds) *Handbook of international historical microdata for population research.* Minneapolis: Minnesota Population Center. 259–284.

Ruggles, S. Hacker, J.D. and Sobek, M. (1995) ‘Order out of chaos: General design of the integrated public use microdata series.’ *Historical Methods,* 28(1): 33–39.

Ruggles, S. and Menard, R.R. (1995) ‘The Minnesota historical census projects.’ *Historical Methods,* 28(1): 6–10.

Seltzer, W. and Anderson, M. (2000) ‘After Pearl Harbor: The proper role of population data systems in time of war.’ Paper presented at: Annual meeting of the Population Association of America, Los Angeles, March.

Tambay, J-L. and White, P. (2001) ‘Providing greater accessibility to survey data for analysis.’ Paper presented at: Joint ECE/Eurostat work session on statistical data confidentiality, Skopje, March.

Thorogood, D. (1999) ‘Statistical confidentiality at the European level.’ Paper presented at: Joint ECE/Eurostat work session on statistical data confidentiality, Thessaloniki, March.

UNESCO (1997) *The International Standard Classification of Education (ISCED 1997).* Paris.

United Nations Statistics Division (1947–). *Studies of census methods.* Department of Economic and Social Affairs, New York.

United Nations Statistics Division (1998) *Principles and recommendations for population and housing censuses.* Department of Economic and Social Affairs, New York.

United Nations Statistics Division (1990) *International Standard Industrial Classification of All Economic Activities (ISIC-88).* Department of Economic and Social Affairs, New York.

United Nations Statistics Division (2001) ‘Population and housing census dates,’ *www.un.org/depts/unsd/demog/cendate*, April. Department of Economic and Social Affairs, New York.

United Nations Economic Commission for Europe and Statistical Office of the European Communities (1998). *Recommendations for the 2000 Censuses of Population and Housing in the ECE Region.* Statistical Standards and Studies, No. 49. New York and Geneva.

United States Bureau of the Census (1964) *Census of population and housing, 1960 public use sample: one-in-one-thousand sample.* Washington, D.C.

United States Bureau of the Census (2000) ‘2000 round census dates.’ *www.cache.census.gov/ipc/www/cendates/cenall.pdf.*

Wright, C.D. and William C. Hunt, W.C. (1900) *History and growth of the United States census.* Washington, DC: Government Printing Office.