Executive Summary of Project Proposal to the National Institutes of Health: European census microdata harmonization project (IPUMS-Europe)

A vast quantity of raw European census microdata for the period since the 1960s survives in machine-readable form. Most of these data, however, remain inaccessible to researchers. This proposal seeks funding to create harmonized and documented samples of over 50 Western and Eastern European censuses. These data will be made available for scholarly and educational research through a web-based data dissemination system.

This project leverages previous federal investments in social science infrastructure. Grants from the National Institutes of Health and the National Science Foundation have laid the groundwork for the European data series by funding many of the initial costs. Those projects have underwritten the development of data cleaning and sampling procedures, data conversion and dissemination software, and design protocols for data and documentation. We have already made arrangements to obtain raw microdata files, internal documentation, and redistribution agreements for over 60 censuses of 17 European countries with populations totaling one-half billion people. As a result, the new public-use microdata samples for Europe will be highly cost-effective.

The following tasks must be carried out to capitalize on these past investments and make the European data widely available to researchers: draw new samples of each census; reformat and clean the samples; impose confidentiality protections; recode variables into existing harmonized coding systems and develop new coding designs optimized for Europe; allocate missing and inconsistent data values; create a set of consistent constructed variables; develop harmonized English-language documentation; convert all documentation to the Data Documentation Initiative metadata standard; and improve and maintain the web-based data access system.

With over 70 million records spanning a forty-year period, the new database will allow social scientists to make comparisons across European nations during decades of dramatic change. Coupled with data from other IPUMS projects, this information will allow innovative comparative research across time and space. The data series will result in a substantial body of new scientific and policy-relevant research on economic transformation, demographic transition and population aging, international migration, and many other topics.

PERFORMANCE SITE(S) *(organization, city, state)*

Minnesota Population Center
537 Heller Hall
271 19th Avenue South
Minneapolis, MN 55455

Minnesota Population Center
425 20th Avenue South
Minneapolis, MN 55454

KEY PERSONNEL. See instructions. *Use continuation pages as needed* to provide the required information in the format shown below.
Start with Principal Investigator. List all other key personnel in alphabetical order, last name first.

Name	Organization	Role on Project
Robert McCaa	University of Minnesota	Principal Investigator
Steven Ruggles	University of Minnesota	Co-Investigator
Miriam King	University of Minnesota	Co-Investigator
Deborah Levison	University of Minnesota	Co-Investigator
Matthew Sobek	University of Minnesota	Co-Investigator
Trent Alexander	University of Minnesota	Research Associate
Albert Esteve	University of Minnesota	Research Associate

Overview

The goal of this project is not simply to make European census data available; it will also make them usable. Even where census microdata can be obtained, comparison across countries or time periods is challenging because of inconsistencies between datasets and inadequate documentation of comparability problems. Because of this, comparative European research based on pooled census samples is rarely attempted. This project will reduce the barriers to cross-national research by converting census microdata into a uniform format, providing comprehensive documentation, and making the data available without cost to researchers through a web-based access system. The European census microdata series will be fully harmonized with IPUMS-International, thus facilitating comparison across continents as well as across Europe.

We anticipate that the European microdata series will include as many as seventy-one million persons in sixty-two censuses from seventeen countries, and there is potential to include additional censuses from other countries.[1] Austria, Belarus, Bulgaria, Czech Republic, Poland, Portugal, and Russia will provide 5 percent samples; Belgium, Greece, Romania, and Slovenia will provide 10 percent samples; and the Netherlands and the United Kingdom will provide 1 percent samples.[2]

For purposes of planning and design, we must work simultaneously with all these censuses. This will ensure that we accommodate the full range of variation across countries and census years when designing harmonized variable coding systems. During data and documentation processing, however, we will work with batches of four countries at a time. This approach—also used for IPUMS-International and IPUMS Latin-America—allows timely release of samples and avoids the logistical complexity of processing too many censuses simultaneously.

We will process as many batches as possible within the five years of this project. Based on our experience with China, France, Kenya, Vietnam, Mexico, Colombia, and Brazil, we hope to complete work on four batches—sixteen countries and fifty-two datasets—within the time frame of the present project. Depending on the extent of data format and consistency problems we encounter, however, that number could change. In addition, we will include in the data series the five French censuses we have already processed and a new sample for 1999. We have established a priority sequence based on intellectual salience, census quality, technical characteristics, and the release schedule for the 2000 round of census data. The proposed processing sequence is as follows:

2005: Belgium, Hungary, Spain, United Kingdom
2006: Austria, Belarus, Bulgaria, Romania
2007: Czech Republic, Germany, Poland, Slovenia
2008: Greece, Netherlands, Portugal, Russia

Foreign Expenses

We have negotiated licenses and fees for the dissemination of microdata for each European country with which we have reached agreements (see attached copy). In year one of the grant, one-half of the total fee for each country will be paid to license the complete set of microdata and documentation for all census rounds and microcensuses before 2000. The remainder will be paid upon receipt of microdata and metadata for the 2000-round census for each country. This is very cost-effective, since the fees cover the cost of supplying the data, and where necessary, translating essential source documentation, and technical support by national experts

[1] These figures include ten microcensuses, which are high-density census-style surveys used as a substitute for census data. In conception and execution the European microcensuses closely parallel the American Community Survey, which is scheduled to replace the long-form of the U.S. census in 2010 and thereafter. In Europe, microcensuses have been used to reduce the gap between complete censuses (Poland, Germany, Russia) or where continuous registration systems have replaced conventional censuses (Netherlands).

[2] We are awaiting final approval from the authorities in Belgium, Poland and Russia, and we are still negotiating the license fee with Germany, but we expect that these agreements will be finalized soon. In addition, there are several earlier censuses for which the readability of old census tapes cannot be confirmed until the license agreements are executed, including Austria 1961, Belarus 1989, Poland 1960, and Slovenia 1981 and 1991.

(e.g., drawing samples from the complete data as needed, supplying necessary documentation, and answering questions). Fees are less for Western European countries, where the agencies have agreed to subsidize the project, and greater for Southern and Eastern European countries. Some of the latter (Greece, Poland and Russia) have several complex datasets for which data must be recovered and samples drawn from scratch. Data for three countries—France (6 censuses), Spain (3 censuses), and Hungary (4 censuses)—are available at no cost, since the license fees were covered by NSF grant SBR 9907416. We are awaiting final approval from the authorities in Belgium and Russia, and we are still negotiating the license fee with Germany, but we expect these agreements to be finalized soon. In addition, there are several earlier censuses for which the readability of old census tapes cannot be confirmed until the license agreements are executed, including Austria 1961, Belarus 1989, Poland 1960, Russia 1989, and Slovenia 1981 and 1991. If any of the datasets do not materialize, we will use the savings to negotiate licenses for censuses from other countries. We also have budgeted a substantial sum in the final project year to cover the costs of additional censuses that become available during the course of the project.

Table 1. Available Censuses and Expected Sample Sizes (in 000s), by Country

	Sample Density (%)	Census	N	Census	N	Census	N	Census	N	Census	N	Census	N
Austria	5	2001	(405)	1991	(390)	1981	(380)	1971	(375)	1961	(360)	.	.
Belarus	5	1999	(520)	1989	(510)	.	.	.	.	.	.	.	.
Belgium	10	2001	(1,030)	1991	(1,000)	1981	(990)	1971	(970)	.	.	.	.
Bulgaria	5	2001	(395)	1992	(425)	.	.	.	.	.	.	.	.
Czech Rep.	5	2001	(515)	1991	(515)	.	.	.	.	.	.	.	.
France	5	1999	(3,005)	1990	(2,361)	1982	(2,714)	1975	(2,629)	1968	(2,488)	1962	(2,321)
Germany	.5	2000	(414)	1991	(400)	1987	(306)	1982	(308)	1970	(305)	1962	(285)
Greece	10	2001	1090)	1991	(1,020)	1981	(970)	1971	(880)	.	.	.	.
Hungary	5	2001	(505)	1990	(520)	1980	(535)	1970	(515)	.	.	.	.
Netherlands	1	2001	(160)	1991	(150)	1981	(141)	1971	(130)	1960	(115)	.	.
Poland	5	2002	(1,930)	1995	(1,940)	1988	(1,900)	1984	(1,850)	1978	(1,745)	1974	(1,680)
Portugal	5	2001	(500)	1991	(495)	1981	(490)	.	.	.	.	.	.
Romania	10	2002	(2,240)	1992	(2,280)	.	.	.	.	.	.	.	.
Russia	5	2002	(7,200)	1989	(7,400)	.	.	.	.	.	.	.	.
Slovenia	10	2001	(200)	1991	(200)	1981	(180)	.	.	.	.	.	.
Spain	5	2001	(2,040)	1991	(1,940)	1981	(1,875)	.	.	.	.	.	.
UK	1	2001	(600)	1991	(574)	.	.	.	.	.	.	.	.

Total person records = 71,101

Notes:

Microcensuses: Germany 1982, 1991, 2000; Netherlands 1981, 1991, 2001; Poland 1974, 1984, 1995.

Our present agreement calls for a .5% sample for France 1999. We hope to expand that sample in the course of the project.

A sample for 1960 Poland is included in the total case count.

Final agreements for Belgium, Germany, Poland and Russia are pending, and some of the earliest censuses may not be recoverable (see note 2).