HISTORICAL METHODS, Spring 2003, Volume 36, Number 2

Occupational Classification in the North Atlantic Population Project

EVAN ROBERTS
Minnesota Population Center
University of Minnesota

LISA Y. DILLON
Département de Démographie
Université de Montréal

MATTHEW WOOLLARD
History Data Service
University of Essex

GUNNAR THORVALDSEN
Norwegian Historical Data Centre
University of Tromsø

CHAD RONNANDER
Minnesota Population Center
University of Minnesota

Abstract. The North Atlantic Population Project (NAPP) is a complete-count data set of late-nineteenth-century censuses from Canada, Great Britain, Iceland, Norway, and the United States. One of the project's most challenging tasks is the coding and classification of 2 million distinct responses to occupational questions. Using the Historical International Standard Classification of Occupations (HISCO) as the basis for their classification scheme, the authors have adapted it to address particular problems applicable to the NAPP occupational data—the inconsistent specification of tasks, industry, and employment status by census respondents; variation among the NAPP countries in the level of occupational detail provided; and spatial and temporal variation in the language used to describe occupations. Compared with HISCO's classification scheme, the NAPP system reduces the overall number of codes, introduces new codes, and retains more detail from vaguely specified occupations.

Keywords: census, Historical International Standard Classification of Occupations (HISCO), microdata, North Atlantic Population Project (NAPP), occupations

Occupation is the only item consistently collected across time and national boundaries that provides comparable information on economic and social status in late-nineteenth-century censuses. The North Atlantic Population Project (NAPP) will gather into a single database the nearly 90 million person records that comprise seven censuses of Canada, Great Britain, Iceland, Norway, and the United States (see companion article by Roberts et al., pp. 80–88 in Part Two of this issue). Adapting and applying a universal occupation classification scheme is one of the project's most complicated tasks, and it is worth examining that process in detail.

Coding Issues

Information on occupation was transcribed for each individual as it appeared on the original enumeration form. The NAPP data set will contain well over 2 million distinct alphabetic occupational strings (unique orderings of alphanumeric characters and spaces—"black smith" and "blacksmith" are distinct strings because of the space in the first string).[1] Fortunately, some occupations were very common, and thus a large percentage of the population is described by a very small percentage of strings. For English-language occupations, figure 1 reports the proportion of the population encompassed by a given percentage of the occupational strings in the data. The most common 1 percent of occupational strings provide codes for more than 80 percent of the working population. Most of the work, therefore, will involve coding occupations for the remaining 20 percent of the population.

Occupational information has value only to quantitative analysts insofar as the multiplicity of reported occupations can be distilled into a concise, meaningful classification system. The nature of the manuscript responses further complicates matters. Open-ended responses to census inquiries about occupation often blur the boundaries between information on employment status, industry, and specific occupational tasks. These problems are magnified when one considers the culturally specific context of the responses; it is a particular challenge to maintain a consistent classification across countries with procedural and linguistic variations.

Occupational terminology is multidimensional. There are three generally accepted dimensions of job descriptions: type of work, status, and industry. Essentially, a person's occupation is what he or she does. The tasks, functions, and skills necessary to carry out that job define an occupation. An industry consists of all the firms—ranging from large companies with thousands of employees to self-employed individuals—that concentrate on producing or selling the same commodity. Within a firm, there is a further distinction between owners, supervisory employees and manage-

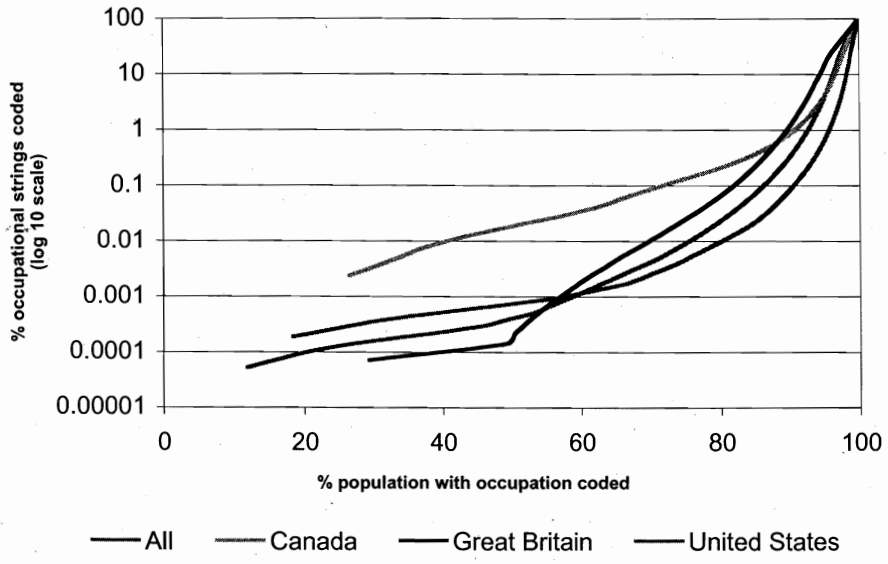

FIGURE 1. Percentage of individuals with coded occupations, by percentage of occupational strings coded.

ment, and those people employed to carry out the tasks. Many firms are so small that the owners employ no assistants. Thus, there are three classes of workers: employers, self-employed people working on their own account, and people working for salary or wages.

Responses in the NAPP database often encompass some description of tasks, industry, and class of worker, but this information has not been recorded consistently. The problem arises because censuses included open-ended questions about occupation that did not require respondents to distinguish between the tasks they carried out, where they worked, and their relationship to an employer. Twentieth-century censuses and occupational classification schemes do distinguish between these three variables, and for this reason there are some difficulties in applying twentieth-century schemes to nineteenth-century data. Moreover, occupational terminology takes on full meaning relative to other contexts and information about the person performing the job. The meaning—even within one language—of an apparently straightforward occupational title can vary depending on the geographic, temporal, or industrial context.

A single occupation is composed of a set of tasks and roles within an organization. Even if information about industry and class of worker is recorded in a separate variable, an occupational classification scheme must reduce the multiple tasks carried out by a worker into one dimension. For example, a foreman in a carpenter shop has a supervisory (foreman) and production (carpentry) role. A foreman in a blacksmith shop shares the supervisory role but not the production role. Some researchers would group these workers together because they both supervise employees, but others would separate them because they worked at different production tasks. An ideal occupational classification scheme would retain information about both the supervisory and production roles, so the individuals could be grouped or separated depending on the research question.

Many occupational responses are too ambiguous to be interpreted reliably. Frequently, a term refers to a different set of tasks in different industries, even though the industry is not specified. A clerk in a store, for example, is not the same as a clerk in a courtroom, though they are more alike than a fireman on a steamboat and a fireman in a fire department. A housekeeper in a hotel and a housekeeper who is working in her own home may carry out many of the same tasks, such as ironing clothes and cleaning. Both may live on the premises where they work. Despite their doing similar work, the hotel housekeeper is generally the only one of the two who has a formal employment relationship and receives wages.

A worker's relationship to his or her employer, often called "class of worker," provides important additional information about a job, especially whether an individual has management responsibilities. All censuses included in the NAPP database provide inconsistent information on class of worker. Apprentices and proprietors often identified themselves as such to enumerators, yet it is certain that many did not. Similarly, as independent artisan trades were mechanized in the nineteenth century, occupations such as "watchmaker," "bootmaker," and "blacksmith" could have been performed by artisans who were independent, who employed a small number of hands, or who worked inside factories.

Enumerators did not always record information they thought would be obvious to the examiner who looked at the completed form. For the NAPP project, we assign codes

in the absence of contextual information available on the manuscript. To consult the manuscripts for ambiguous cases would be prohibitively time consuming and expensive. This necessary labor-saving practice affects our knowledge of tasks, industry, and class of worker. For example, hotelkeepers were sometimes just recorded as "keeper." Without the contextual information, "keeper" does not clearly specify the tasks the respondent performs. Similarly, the manuscript would make it clear that a "supervisor" in a household that was an almshouse was an almshouse keeper. As they carry out their work, the only word NAPP researchers see is "supervisor," which would receive a code for foreman and is likely to be grouped near manufacturing foremen, not with service workers.[2]

National and Temporal Variation in Occupational Responses

The level of occupational detail found in the census varies significantly across countries. In the United States in 1880, for example, respondents often identified the industry in which they worked, without identifying the tasks they performed. Thus, it was common to see vague responses such as "works in cotton mill." These workers could be classified as similar to others in the same factory who did specify their tasks (e.g., spinners). But the vague responses could also be grouped together so that "works in railway shop" would be coded with "works in cotton mill." Outside the United States, this problem is rare; for example, there were only 108 responses in the entire Canadian census of 1881 that include "works in," "works at," or "works on."

The variation in occupational detail provided by each national census resulted partly from differences in enumeration instructions and practices. In general, the most detail about occupations was obtained by British enumerators, who were instructed to ask about the numbers of employees and the acreage of farms. Such information often helps to establish the class of worker, and it usually allows farmers who employed laborers to be distinguished from those who farmed by themselves.

Differences in occupational information attributable to enumeration practices are sometimes compounded by linguistic variation. Although harmonizing data from five countries with four languages intensifies some of the aforementioned classification challenges, these problems also exist within the same language or even the same country. Our experience with usage in the English language uncovered three types of linguistic problems: the same word refers to a different set of tasks in different places, different words describe the same occupation in different places, and words used to refer to an occupation changed over time.

The best example of the same word referring to different tasks is "engineer." In the United States, an "engineer" could be a stationary engine operator, a railway locomotive driver, or a professional civil or mechanical engineer. We expect that most professional engineers identified themselves by their full title, so we classify "engineers" as engine operators. In contrast, someone reporting "engineer" in Great Britain was more likely to be a professional, though some were simply managers of machinery, as in the United States. We classify "engineers" in Great Britain as professional engineers. Although these jobs are not completely unrelated, the marked differences in education and supervisory responsibilities require that engine operators and professional engineers should be given distinct codes.

Conversely, some occupations are described by different words in different countries. We must acquire knowledge of the tasks that make up these jobs to decide which terms refer to work sufficiently similar so that they can be classified together. For example, British "chemists" and American "druggists" should be assigned the same code because they do the same work. Even within a given country, some occupations are described by several terms that do not mark any important distinctions in tasks, industry, or class of worker. In the United States, "merchant," "storekeeper," and "dealer" are synonyms. In Canada, "chemist" embodied the English meaning, yet many Canadians were also using the U.S. title "druggist." These titles describe the same occupation and should receive the same code.

Occupational terminology is also fluid over time. NAPP data are drawn from countries in various stages of industrialization, and within those countries the process of industrialization had proceeded at different rates. Because NAPP will form the basis for an extended time-series of harmonized data, the occupational classification must be sensitive to changes in the relationship between an occupational term and the job it describes. With industrialization, for example, occupations such as "shoemaker" shifted from self-employed craftwork to wage employment in a factory setting. Professionalization, as well as industrialization, changed the way in which occupations were performed and described. For example, "nurse" now often refers to a post-secondary-trained worker based in a hospital, but in 1880 the word was in transition. In the United States at that time, "nurse" generally referred to a woman who cared for sick people or children in a private household. In the United States, professional training of nurses began in the 1870s, so by 1880 some "nurses" were hospital trained and employed (Reverby 1987). The training and employment relationships of the two groups differ substantially.

In addition to English titles, we must classify occupations in French, Norwegian, and Icelandic. We use English for documenting the classification system because it is the language of most of the source material and of most coding schemes designed for cross-national use. However, some terms do not have a clear English-language equivalent. It is necessary, for example, to distinguish a "*habitant*" (Canada) and "*feuar*" (Scotland) from a "farmer."

Idiosyncrasies of each country require special treatment. In Norway and Iceland, the titles of "farmer and fisherman"

and "*cottar* and fisherman" are among the most common occupations (a *cottar* is a small-scale farmer). Instead of classifying these cases as either farmers, *cottars*, or fisherman—which would lose important information about the nature of their work—we created new titles that reflect the combined occupations. We also retain nonstandard information specific to each country's enumeration instructions. For example, the use of the term "farmer's son" was specifically delineated in enumerators' instructions in Great Britain.

Many of the common occupational strings have substantial overlap among the three primarily English-speaking countries. In Great Britain, Canada, and the United States, just 6,601 common strings account for between a quarter (Canada) and a half (United States) of each population's occupations. If we add in occupations common to at least two of the three countries, we arrive at a total of 44,175 occupational strings, which cover the great majority of the population in Canada and the United States and half the population in Great Britain (figure 2). In the overlap lie both the challenge and the opportunity. Collaboration and crosschecking will ensure that the data coding is of a higher standard than the individual country teams would achieve on their own.

We recognize that our judgments concerning which distinctions and details are important may not be shared by others. Accordingly, researchers will be able to download the full alphabetic occupational strings as well as numerically coded data. Users can then reclassify occupations for their own analyses and evaluate our coding logic.

Classification Scheme

The goal of the NAPP occupational classification system is to provide users with an international system containing a manageable number of categories without losing significant information from the original string. We also needed a system that was easy to learn and apply to our data sets; with more than 2 million strings to code, we must be very efficient. Although we were familiar with national occupational classification systems used for some individual databases, we determined that none of these systems would provide a suitable scheme for an international data set (Sobek and Dillon 1995; Ronnander 1999; Thorvaldsen 1995; Woollard 1999).[3] The NAPP project employs a modified version of the Historical International Standard Classification of Occupations (HISCO), a new classification scheme created for historical international data (van Leeuwen, Maas, and Miles 2002).

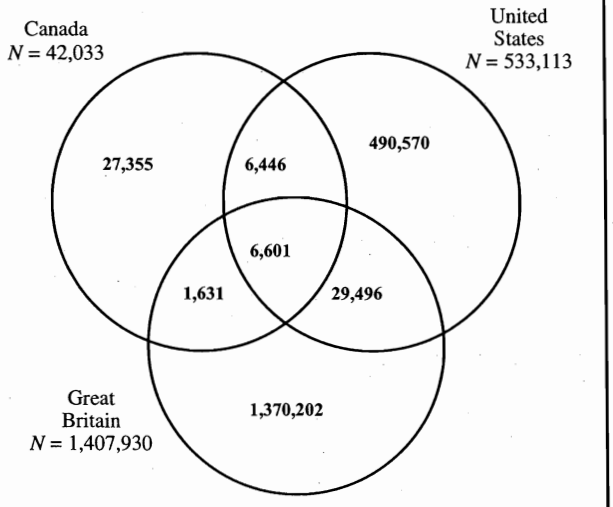

FIGURE 2. Correspondence of occupational titles in three of the NAPP data sets.

The HISCO scheme classifies historical occupational data designed for use across a long time span and across multiple countries. We based our codes on HISCO because we expect it to become an international standard for classifying historical occupational data. The HISCO system is based on the 1968 version of the International Labour Organisation [then known as Office]'s International Standard Classification of Occupations (ISCO68) (ILO 1969). HISCO was developed by a team of researchers from Belgium, Canada, France, Germany, Great Britain, the Netherlands, Norway, and Sweden. These investigators used occupational data derived from a variety of historical sources, including birth, death, and marriage certificates, census records, and parish, civil, and catechetical registers to produce an index of occupational titles and inform a modified version of ISCO68.

The HISCO system required some modifications to make it practical for the massive project of coding the 2 million titles in the NAPP database. During November 2001, representatives from all of the NAPP country research teams met in Minneapolis for 10 days to prepare a modified coding scheme. We worked primarily with the HISCO system but also drew on the scheme for the 1881 U.K. census, ISCO68, the U.S. Census Bureau's scheme from 1950, and the Norwegian system from 1900.

There are two significant differences between HISCO and the NAPP coding scheme. First, the NAPP scheme has far fewer codes (about 650, as compared with 1,881). Many eliminated titles do not appear in the historical data and therefore are not needed for the NAPP project. Second, we have added some new categories and moved selected occupations between groups. These revisions resulted in mostly subtle differences, but in a few cases there were significant changes. The main intentions of our adaptations were to

- reduce the number of groups in use to facilitate mass coding;
- introduce new groups that were of specific importance in our respective national data sets and create new groups for vague and imprecise occupations;
- create additional documentation and indexing to allow users of the NAPP data set to understand the coding decision-making process;

- retain as much compatibility with HISCO as possible by making translation tables available and by trying to make each HISCO code map to just one NAPP code.

In working with both the existing HISCO classification scheme and the occupational responses from the existing complete-count data sets, our classification is driven by empirical observation as much as by theory. The most commonly occurring terms receive their own heading and code. In naming headings and unit groups, we have followed the terminology used by respondents.

We have tried to make the NAPP occupational codes compatible with HISCO's whenever feasible. There are, however, differences between them that preclude a complete correspondence between codes. As described in the following sections, we have deleted some HISCO codes and added others. In a few cases, we also rearranged the headings. For example, the heading Farm Manager, which HISCO classifies in major group 2 (Administrative and Managerial Workers) has been moved to major group 6 (Agricultural Workers).[4] We will make available a variable that classifies occupations in the NAPP data set into HISCO major groups. Our documentation will allow users to optimize comparisons between our classification and the original HISCO codes. We plan to provide full instructions for recoding both the HISCO and NAPP classifications to make them as compatible as possible.

Code Reduction

Like ISCO68, HISCO is based on a four-level hierarchical code. These codes consist of a major group, a minor group, a unit group, and a heading. For example, the heading "Hotel Receptionist" (3-94.20) is found in the unit group "Receptionists and Travel Agency Clerks" (3-94), in a minor group "Clerical and Related Workers Not Elsewhere Classified" (3-9), in the major group "Clerical and Related Workers" (3). HISCO has 8 major groups, 83 minor groups, 284 unit groups, and 1,881 headings.

Within HISCO, a large number of unit groups make fine distinctions among headings that are inapplicable or indeterminate in late-nineteenth-century census data. In many instances, the exhaustive classification of HISCO suggests false precision when applied to nineteenth-century censuses. For example, there are 12 headings within the unit group "Shoe Cutters, Lasters, Sewers and Related Workers," and there is another more general unit group "Shoe Makers and Shoe Repairers." The occupational responses in the NAPP data sets generally have insufficient detail to make use of such fine distinctions. The HISCO unit group "Railway Engine-Drivers and Firemen" (9-83) includes the following headings:

9-83.00	Railway Engine-Driver or Fireman, Specialisation Unknown
9-83.20	Railway Engine-Driver
9-83.30	Railway Steam-Engine Fireman
9-83.40	Assistant Railway Engine-Driver
9-83.50	Underground or Elevated Train Driver
9-83.60	Engine-Driver (Mine or Quarry)
9-83.90	Other Railway Engine-Drivers and Firemen

This degree of specificity is rarely found in late-nineteenth-century censuses; men who drove or ran trains would often describe their occupation as "driver on R.R.," "drives train," "engine-driver," or "runs train." These responses do not indicate whether the driver was an assistant, drove a steam engine, or worked on a particular railway. Thus, in the NAPP coding scheme, we have simplified these codes to distinguish between those specifically referred to as drivers of engines and those who were explicitly "stokers" (U.K.) or "firemen" (U.S.).

In some instances, the detail in HISCO is inappropriate because nineteenth-century workers were less specialized than they are today. Thus, someone describing himself as a watchmaker in the nineteenth century may have been carrying out many of the subdivisions of this trade defined by HISCO.[5] Under the influence of scientific management, the subdivision of jobs in many industries has led to more narrowly defined occupational roles than in the late nineteenth century.

We were able to collapse many HISCO headings into new aggregated headings with minimal loss of information; we retained all categories with a significant number of responses. This simplification of the coding system helped us to develop the cost-effective procedures necessary for such a large coding project. The reduction of headings addressed our need for a system that coders could learn quickly. It is obviously much easier for coders to familiarize themselves with 650 codes than with the 1,881 headings in HISCO.

New Codes

Several important occupational and industrial groups do not appear in the HISCO classification. Given the size of the complete-count NAPP database, it is unsurprising that we encountered occupations that the HISCO project had not. Because the English-language occupational titles in HISCO came from marriage registers, we omitted some historical occupations carried out mainly by the young and the aged. We have therefore introduced headings for errand- and office boys/girls and have paid greater attention to those outside the labor market who gave a nonoccupational response—for example, "annuitant," "scholar," and "unemployed."

We also introduced new headings for late-nineteenth-century occupational responses that have since generally disappeared. In the chemical industry, where HISCO (following ISCO) had based occupational headings on processes (e.g., crushers, cookers, or roasters), we introduced product-specific groups (e.g., drugs, gunpowder, or charcoal). Most nineteenth-century chemical workers did not report their specific tasks but instead reported what they produced.

In general, we added categories based on a rough frequency threshold. For example, we added headings for toll collectors and clog makers because we encountered a significant number of cases with these titles. Conversely, although professional musicians sometimes stated the instrument they played, we did not create separate codes for "cellists" and "pianists," because they occurred infrequently.

Reflecting our preference for following the terminology used by respondents, we have added new codes for such terms as "machinist," "blacksmith," and "gardener." For many of these common occupational titles, such as "machinist," there is no single obvious code in HISCO; as a result, different coders could legitimately place the same occupation in several different places. For example, "railwayman" could be given any one of eight different codes in HISCO, in the managerial, clerical, and transport operator major groups. In the absence of any clarifying information, we thought it was potentially misleading to allow the same string to be coded in so many different ways.

Although these general titles may have varying meanings across workplaces, time, and nations, we cannot infer these meanings from the responses. For example, in HISCO "Blacksmith, General" was defined as "forges and repairs articles of iron and steel, such as hand tools, hooks, chains, agricultural implements and metal structural parts, using hand or power hammers." Included within this heading were some occupational titles that demanded different skills and levels of experience, such as "anchorsmiths" and "farriers." We created a new heading for people described simply as "blacksmith" and additional new headings for the more specific titles when they were numerically significant.

We also introduced several modifications to accommodate the vague occupational titles often found in census data. The most common of these is the "works in [specified type of] factory" response. HISCO offers two ways of coding such cases. The preferred method is to place these workers in a residual "factory worker" heading (9-99.30). Thus, these workers are separated from the bulk of titled and semiskilled workers in the same industry who may be kindred workers who just happened to have reported their work more accurately. The second method is to integrate these responses with more explicit ones in the same industry. This practice, however, implies that we know the workers are semiskilled operatives when the data are simply unclear. Because of the scale of this problem, we have assigned particular codes to workers who gave "works in [specified type of] factory" responses. In turn, we moved laborers in specified industries to the same unit group (three-digit codes) as the titled operatives in the same industry, to prevent a false distinction between "laborers" and "works in" responses.

In HISCO, laborers are relegated to their own general code (9-99.10), which does not preserve industrial distinctions. Our modified classification allocates manufacturing workers for a particular industry to their own unit group(s) that contains skilled and semiskilled workers, "works in" respondents, and laborers. Differences among the three groups are preserved in a separate two-digit variable termed *status*. This approach allowed us to accommodate variations in the level of detail about tasks and industry. It also allows users to identify some manufacturing industries without using a separate variable.

We made a variety of additional modifications to handle other kinds of vague responses. For example, a common English-language title for managers and proprietors in retail or wholesale trade was "store/shopkeeper," and the product being sold was often, but not always, specified. It is not clear from this title whether the individual was an employee or the storeowner. In the HISCO classification scheme, a "storekeeper" could be coded into two separate codes for either employed sales managers or "working proprietors." Because the data do not support the distinction, the NAPP system groups these two categories together under one code: "working proprietors in wholesale and retail trade."

Finally, as noted earlier, we added several codes to retain important country-specific occupational information. These include the job titles "farmer and fisherman" and "*cottar* and fisherman," which were common in Norway and Iceland, "teamster" a U.S./Canadian term only, "*habitant*" specifically for Canada, and "farmer's son" for Canada and Great Britain. We have tentatively included subheadings (distinguishable at the fifth digit) for jobs when we can differentiate between the type of contract held. This modification was made specifically for titles in Britain that could be carried out by domestic servants or those in the commercial sector, including occupations such as groom or coachman.

Because we are classifying such a large number of occupations, our additions to HISCO may prove useful to other researchers wishing to code nineteenth-century data. England's relatively advanced industrialization in 1881 ensures that almost all occupations occurring in North America or the rest of Europe in that period will have a place within our scheme. Furthermore, wide variation in the specificity of occupational information provided by the censuses forced us to create a scheme capable of handling data ranging from detailed to general, while retaining comparability between data sets.

Occupational Descriptions

We have modified many occupational descriptions to better fit nineteenth-century work. Some ahistorical categories, such as airline pilots, are innocuous. Obviously, there were no airline pilots in 1880, and the availability of this code is not problematic, because the code does not contain other occupations that did exist in 1880. Such headings are not relevant to the NAPP project, but we have retained space for them to ensure compatibility with more recent data.

In some cases, we altered occupational descriptions to ensure that nineteenth-century occupations are correctly classified. For example, the term *hairdresser* is described in

the HISCO manual as: "cuts, washes, tints and waves hair and performs other personal services incidental to women's hairdressing." A contemporary dictionary of trade terms describes hairdresser as: "an artist who trims and arranges the hair; a *perruquier* [wigmaker], who often combines the sale of perfumery and toilet articles" (Simmonds 1858). The first part of the contemporary definition—regarding cutting and styling hair—accords with HISCO, but the second part about wigmakers is distinctly different. In that case, we have altered the definition of hairdresser in our documentation, and the term "*perruquier*" would be classified as "wigmaker," not "hairdresser."

Even when classification was not affected, we modified some job descriptions to include the tasks and occupational responses we encountered in the complete-count nineteenth-century census data. For example, the HISCO description of *firefighters* reads:

> Workers in this unit group extinguish fires, eliminate fire hazards and protect property at fire sites. Their functions include: fighting fires as members of a public or private fire-fighting force; detecting and eliminating or reducing fire hazards in industrial plants or other establishments; protecting and salvaging goods during and after fires; preventing or extinguishing fires in crashed or damaged aircraft and rescuing crew and passengers; performing other related duties.

By contrast, the NAPP definition is broader:

> Workers in this unit group include all regular members of public or private fire-fighting forces charged with preventing, extinguishing, and containing fires and protecting people at fire sites. This includes overseers of fire-fighting units and specialized members of fire-fighting forces such as fire engine drivers, pumpers, and hose men, but not employees such as clerks, mechanics, or fire-house keepers, whose tasks require no knowledge of fire-fighting.

The new description includes people doing the very specific tasks in the original description but encompasses the general and vague responses encountered in late-nineteenth-century census data.

Subsidiary Information

Occupational responses often contain relevant information that cannot be incorporated within a classification scheme. To preserve this information for researchers, we have followed the HISCO framework for retaining three categories of data: the product a person was making or selling, a worker's *status* within the workplace or labor market, and family relationships.

In manufacturing and trade industries, occupational data often contain information on the product being made or sold. Following the HISCO system we retain this information in a separate *product* variable, classified according to the three-digit *group* level of the UN Central Product Classification (CPC) Version 1.1. One of the main advantages of this classification system is that it provides for vague, general, and ambiguous responses—for example, "general shopkeeper" as well as more specific terms.[6] Manufacturing activities are often stated more specifically than distribution activities. For example, many more people report the occupation "shirt maker" than report working in a "shirt store." There are, however, many "clothing stores/shops."

We have also retained the HISCO framework for handling information on status when it is found in occupational data. For example, a person may respond that he is a "retired brick maker." Although that person is no longer performing the occupation, we still provide a code for "brick maker." We then indicate in a separate variable that the person is "retired." Similarly, people who report that they are "unemployed" but have an occupation will receive both an occupation and status code. This approach allows us to retain and summarize more information for users. These data, however, were not collected consistently and cannot be used to generate accurate statistics on unemployment or retirement rates. Some people who were retired or unemployed may have given an occupational response to the census enumerators without also noting that they were out of work. Conversely, some retired or unemployed individuals may have reported other nonoccupational responses. The status variable will also include references to skill level and class of worker, such as masters, journeymen, and apprentices. Status can also incorporate information about nobility and prestige titles given as an occupational response.

Occupational responses often included information on family relationships, such as "blacksmith's wife," "physician's daughter," "banker's son," or "seaman's nephew." We will retain this familial information in a separate relationship variable (distinct from relationship to household head). For an occupational response such as "blacksmith's wife" we will code the occupation under a residual, nonoccupational heading but indicate in the relationship code that the respondent was a wife. We hope that the availability of this variable will spur researchers to analyze the importance of this common form of nonoccupational response.

Implementing the System

Our aim is to provide a comprehensive list of occupational responses with their appropriate headings and codes. A classification system intended for large data sets is not useful unless researchers can apply the codes decisively, rapidly, and consistently. If coders must make too many hurried decisions, they will never complete their work. Some level of individual decision-making cannot be avoided—indeed, it is a requirement—but clear documentation and comprehensive indexing reduce coding time and maximize comparability within the data set.

We have several coders from different countries, and the cross-country comparability of the data set will be impaired if coders make conflicting interpretations of similar terms. Such differences could create an egregious

problem if they affect large occupational groups. The large number of strings precludes our having coders from two countries review every string. To maximize cross-national compatibility, we have examined the thousand most common strings in each country and have verified that researchers from every country knew how to code them. This initial work of comparing and discussing codes for the most frequent strings has been vital to maintaining consistency. In addition, we are exchanging random samples of strings unique to one country for verification by coders from a second country. This exchange helps to ensure that the classification scheme is being interpreted consistently and that the rate of coding errors is kept to an acceptably low level.

The unique NAPP occupational strings are stored in a Microsoft Access database. To code the data, a researcher views small groups of similar occupations, based on a common word or group of letters, such as "fmr" (a common abbreviation for "farmer"), "clerk," or "dry goods." Coders individually examine each string and either manually assign codes to specific responses or use SQL Update statements to code entire subsets. In general, our coding philosophy is that if two occupations are stated, we code the first one (Woollard 2001). Coders, however, can also use their own judgment if the second job title clarifies the first response. For example "broker and real estate agent" would be coded as a real estate agent instead of a stockbroker. By contrast, "farmer and member of legislature" is coded as a farmer, because these are two distinct jobs.

We pay particular attention to checking the strings that appear in two or three countries. We code identical strings consistently unless we have clear evidence that the terms had different meanings in different countries. Figure 2 shows the distribution of English-language occupations across Great Britain, Canada, and the United States and the duplication between two or three of the countries. Many of the English-language responses that occur in only one country are simple variations on responses given in two or three of the countries.

We will release a preliminary version of the NAPP database containing data from all countries except Iceland in summer 2003. It will contain occupations coded to the Norwegian, Canadian, British, and U.S. domestic coding schemes. Harmonized occupation codes will be available for approximately 95 percent of the population in each country.

NOTES

1. At present, the dictionary is only complete for the Church of Jesus Christ of Latter-day Saints (LDS) countries (Britain, Canada, and the United States), with a sample of the Norwegian occupations.
2. We plan a second phase of coding for occupations affected by institutional residence and the relationship of an individual to the head of household.
3. For each country that has been coded to a national classification scheme, occupational data will also be available to users. Researchers for British, U.S., and Norwegian projects have already classified occupations in the NAPP database according to national systems. The 1881 censuses of England, Wales, and Scotland are coded in the original 1881 British system. The Minnesota Population Center has applied the U.S. 1950 Census Bureau classification system to the 1880 U.S. data, as it has done for other samples it has created. The Norwegian Historical Data Centre has encoded the main occupations in the 1900 census according to a two-dimensional system: hierarchical position and economic sector (Sobek and Dillon 1995; Ronnander 1999; Thorvaldsen 1995; Woollard 1999).
4. We believe that in the late nineteenth century farmers frequently carried out the same tasks as farm managers, and thus the two responses should have adjacent codes. Furthermore, there is evidence that some farm managers also carried out tasks similar to those of agricultural laborers, so to place farm managers in a managerial role was little more than anachronistic.
5. Dial Silverer, Glass Maker, Balance-wheel Maker, Cap Maker, Case Gilder, Case Maker, Chain Maker, Cock and Potence Maker, Dial-Plate Maker, Enameller, Escapement Maker, Finisher, Fitter-In, Frame Mounter, Fusee Maker, Gilder, Hand Maker, Joint Finisher, Key Maker, Motion Maker, Movement Maker, Pallet Maker, Pendant Maker, Pinion Maker, Pillar Maker, Screw Maker, Secret Springer, Spring-Liner, Spring Maker, Stop-Stud Maker, Wheel Cutter . . . are most of the different titles involved here. (Listed in Simmonds 1858.)
6. The classification system is available from http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=16 [retrieved: 20 August 2002].

REFERENCES

International Labour Office (ILO). 1969. *International standard classification of occupations. Rev. ed. 1968.* Geneva: International Labour Office.

Reverby, S. 1987. *Ordered to care: The dilemma of American nursing.* Cambridge: Cambridge University Press.

Ronnander, C. 1999. The classification of work: Applying 1950 census occupation and industry codes to 1920 responses. *Historical Methods* 32: 151–55.

Simmonds, P. L. 1858. *A dictionary of trade products, commercial, manufacturing and technical terms.* London: G. Routledge.

Sobek, M., and L. Dillon. 1995. Interpreting work: Classifying occupations in the public use microdata samples. *Historical Methods* 28: 70–73.

Thorvaldsen, G. 1995. The encoding of highly structured historical sources. *Computers and the Humanities* 28: 301–5.

Van Leeuwen, M. H. D., I. Maas, and A. Miles. 2002. *HISCO: Historical International Standard Classification of Occupations.* Leuven, Belgium: Leuven University Press.

Woollard, M. 1999. *The classification of occupations in the 1881 census of England and Wales.* Colchester: University of Essex.

———. 2001. *The classification of multiple occupational titles in the 1881 census of England and Wales.* Colchester: Working paper IV, Department of History, University of Essex.