


# A New Occupational/Industrial Coding System for 19th Century U.S. Heavy Industrial Workers

SAGE Open  
October-December 2015: 1–14  
© The Author(s) 2015  
DOI: 10.1177/2158244015621116  
sgo.sagepub.com  


Richard Healey<sup>1</sup>

## Abstract

Many census occupational classification systems have been developed over the last 150 years. Availability of digital census data sets now means such classifications can be systematically analyzed. Examination of heavy industrial workers in the full count U.S. 1880 census, and other censuses, has revealed major problems in the attribution of occupations to industrial sectors. This is traceable to the original enumeration process, and it particularly affects generic tradesmen such as blacksmiths and carpenters, who worked in numerous industrial sectors. As a result, the imputation of industrial sector codes from recorded occupations by the North Atlantic Population Project (NAPP) is substantially in error, suggesting that re-coding of existing census records using non-census sources would be necessary for such industrial sector codes to have empirical validity. A new occupational/industrial coding system, incorporating the NAPP-modified HISCO scheme, is presented. This system is capable of supporting both future re-coding work, in a structured data warehouse environment, and the systematic coding of occupational data from a range of archival sources such as company records and city directories.

## Keywords

occupational coding, census analysis, North Atlantic Population Project, industrial workers, HISCO codes

## Introduction

Interest in the problem of classifying and coding occupations can be traced back at least as far as the 1850 population census, and in subsequent decades it became a major focus of attention by census officials on both sides of the Atlantic (Conk, 1978; Woollard, 1999). Reasons for this are not hard to find. Whether it was tracking the overall progress of industrial society, urban industrial specialization, the social mobility of immigrants or finding surrogate measures of wealth and social class, occupational information was one of the most valuable tools available to census statisticians and officials in other government agencies (Edwards, 1933). However, the quest for a satisfactory system of classification for the U.S. census was still in progress at the start of the 20th century (Hunt, 1909).

Much more recently, the growing availability in digital form of large historical population data sets containing individual-level data, anonymized or otherwise, has re-kindled interest within several academic disciplines in the seemingly rather dry topic of occupational classification. However, this same availability of large data collections, which are now open to systematic evaluation using powerful database technologies, in ways that were infeasible until relatively recent times, has begun to raise a number of questions. These relate

not only to the 19th century census enumeration practices and published statistics based on the resulting 19th century census figures but also to the validity, reliability, and “fitness-for-purpose” of electronic coding, both of occupational data transcribed from manuscript census schedules and of other measures derived partly or wholly from these data. It is important that such questions are examined sooner rather than later, as funding bodies are increasingly relying on the availability of large secondary data sets, as part of a drive for efficient use of public monies for research. Yet, in most cases, much less effort has been expended to date in determining the quality of these data sets and their suitability for different types of analyses, than would ideally be the case, given they are intended to form an accepted part of general research infrastructure internationally.

It is important to stress at the outset that this is not primarily a criticism of the leading U.S. and European data archives and centers, such as the member organizations of the North

---

<sup>1</sup>University of Portsmouth, UK

### Corresponding Author:

Richard Healey, Department of Geography, University of Portsmouth, Buckingham Building, Lion Terrace, Portsmouth PO1 3HE, UK.  
Email: richard.healey@port.ac.uk



Atlantic Population Project (NAPP), which have undertaken sterling work in making large demographic data sets accessible to researchers (Minnesota Population Center, 2008). Rather, a distinction needs to be made between four potential sources of problems in these data sets, two of which relate to the original manual collection and processing and two to the much later phase of conversion into digital form. The first derives from inherent shortcomings in the collection of the original data, while the second relates to the methodology used for the subsequent manual classification/tabulation, whose end products were the published census tabulations. The third involves transcribing and data entry errors in the digitization process and the fourth is the digital equivalent of the manual classification problem, namely, how data fields within the digital census records are to be coded accurately and consistently. A further question, which has been examined in the course of several large projects, is how coding consistency can be extended beyond national boundaries to encompass international comparisons, although this topic goes beyond the scope of the present discussion.

As there are numerous data fields, even in late 19th century census records (the digital version of the U.S. 1880 census has about 90 fields, for example), a detailed examination of these four types of problems in relation to each data element in turn would be a major undertaking. The focus of attention here will therefore be restricted to an examination of quality issues surrounding the coding and classification of occupational and related industrial categories, and how they can be addressed, as these are among the key types of information utilized by researchers (e.g., Hirschman & Mogford, 2009; Sarkar, 2009). The topic will be further restricted mainly to consideration of the U.S. 1880 census because it is the only complete count census available for that country in digital form and thus it is becoming widely used as a reference point for all types of historical demographic analysis, even when earlier and later sample census data sets are also used in combination with it (Ruggles et al., 2010; Sobek & Dillon, 1995). That said, questions have been raised about the possible manipulation of occupational data, at the enumeration and processing stages, for the young, the elderly, and married women in this census (Carter & Sutch, 1996). It is also recognized that the 1880 time point was part of an extended process of “learning by doing” in the planning and execution of decennial census-taking, so it must also be set within this broader context.

This article begins by identifying a number of problems inherent in the approach to coding of occupations adopted by NAPP for the 1880 census, and by extension for the earlier and later census samples the latter project has also made available. These problems point to the requirements for a new coding system that removes the limitations identified. The main body of the article explains the design and implementation of this system that provides an operational basis for commencing the long-term and difficult process of re-coding industry sector codes in historical population censuses, which must

necessarily be undertaken using non-census sources. In the process, it will also be made clear that the new system can be used to classify and standardize employment and occupational data from non-census sources independently and additionally to its deployment in support of future work on re-coding of industry sector information in large census data sets.

## **Occupational and Industrial Sector Coding in the 1880 Census**

At the outset, some examples serve to indicate why such a study needs to be undertaken. The first relates to the U.S. railroad sector, a very important contributor to the processes of 19th century industrialization (Chandler, 1965; Vance, 1995). For 1880, there are two independent sources of railroad employment data. The first is the person-level records from the population census, where individuals could identify themselves as working in a railroad-related occupation (U.S. Census Office, 1883a). The second is a quite different special report on transportation, where each railroad company was asked to notify the Census Bureau of the total number of people in its employment (U.S. Census Office, 1883b). Although a few local short-line railroads in isolated areas doubtless escaped enumeration, all lines of any substance could be identified relatively easily, in terms of ensuring quite comprehensive data coverage. As the companies are all named in the report, it is also easy now, as it was then, to check the figures against other reports made to State Railroad Commissioners in the different states. Such checks suggest that considerable confidence can be placed in the reported figures, as state officials with local knowledge would likely have been able to identify any attempts at systematic misrepresentation in the data. The final employment total from the Census special report gives 418,957 workers in the railroad sector.

In the digital 1880 population census from NAPP, the original transcribed text strings describing occupations associated with individual records have been standardized and coded using a variant of the standard Historical International Standard Classification of Occupations (HISCO) scheme (explained further below) into many hundreds of numerical categories represented by the US80A\_OCC variable. A second coding of occupations places them on a 1950 basis using the variable US80A\_OCC50US, to try and provide a consistent classification across multiple censuses, although this particular variable will not be examined here further. Of the US80A\_OCC variable values, 18 codes refer directly to different aspects of steam railroad work and a further three doubtless include railroad employees, but may also overlap with horse-drawn street railroad employment, for example, code 36010—unspecified conductors. Counting the records assigned to the 18 codes across the entire census yields a total of 237,480 workers. Adding in the three less specific categories raises the total to 251,490. The first figure very closely matches that of 236,058 for 1880 given by Edwards

(1943, p. 109) in his classic article on occupational trends in the U.S. census, though there is no documentation on exactly how that specific figure was obtained. This does, however, suggest that the recently developed digital coding system for occupations closely reproduces earlier manual findings. This would support the view that present day transcription and coding has not introduced any significant new sources of error, a similar finding to that reported by Woollard for work on the historical censuses of the United Kingdom (Woollard, 1999). That said, it is apparent that the railroad employment total based on the 18 codes is only 56.7% of the total from the 1880 Special Report and adding the other three codes has relatively little effect. Despite the substantial difference in these two totals, there is no clear evidence from the literature that this rather important disparity has ever been noticed or made the subject of further investigation.

One possible approach to resolving the problem would initially appear to be to use another variable from the 1880 NAPP data set, namely, the industry classification (variable US80A\_IND50US), though this classification is also on a 1950 basis for comparative purposes (cf. Ronnander, 1999). This variable has a code (506) for “railroads and railway express service,” which has already been used in the published literature as part of a comparison of employment changes in industrial sectors over time, although in this case the IPUMS census samples, which use the same industrial codes as the full NAPP data set, were used (Hirschman & Mogford, 2009; Ruggles et al., 2010). The industry code only identifies 266,659 railroad employees in 1880. This is a modest increase over the occupation code count, but it is less apparent what the derivation of this figure is, as it does not correspond to the earlier calculations by Edwards noted above. Although neither the NAPP documentation nor the standard reference on NAPP occupational coding make this clear (Roberts, Woollard, Ronnander, Dillon, & Thorvaldsen, 2003), the industry code is necessarily very largely imputed from the occupation data by the NAPP project and is not an independent and additional source of data on individuals, as there is no column in the manuscript census schedules for industrial sector. The effect of this imputation can be seen by cross-referencing the occupational and industrial codes attached to individual records. Taking an example of five industrial states (Pennsylvania, etc.), which have a total of 61,616 individuals given industry code 506, of these almost 88% have one of the 18 railroad occupation codes, a figure that rises to 91%, if the three less specific categories are included. Working in the other direction, nearly 99% of individuals with one of the 18 occupation codes have an industrial code of 506, or nearly 97% if the wider definition is used. Thus, in the vast majority of cases, the industry code provides no additional information over the occupation code. Judging by detailed examination of the original transcribed occupation text strings (which somewhat negates the value of having a code), the limited number of cases where the industry code does provide additional information reflect

situations where additional non-standard text in the string in question allowed a more precise industrial sector attribution to be made. For example, a worker might be described as “boilermaker in the B&O shops,” which would identify him as a Baltimore and Ohio Railroad employee for the industry code, but under occupation he would be standardized to just “boilermaker.” The data coders have thus endeavored to make maximum use of any data present in the census. Despite this painstaking work, still only 63.7% of the railroad workforce can be identified on an individual basis, leaving in excess of 150,000 workers unaccounted for in this industrial sector alone. Similarly, problematic findings have been reported for the anthracite coal mining sector, a large employer in Pennsylvania, though in this case, use of occupational data gave better results than the industry variable (Healey, 2011).

This earlier study traced the source of the discrepancies in the employment counts to the distinction between workers in generic occupations, such as blacksmiths and machinists, and those in industry-specific occupations, such as coal miners or railroad brakemen. The large numerical impact of these discrepancies does not appear to have been recognized in previous studies devoted to the problem of occupational coding. In general, industry-specific occupations were quite accurately recorded in the population census, so industry sector can usually be imputed correctly for these workers. However, the vast majority of generic workers in 1880 did not give census enumerators details of the industrial sector in which they worked, so it is not possible to impute the industry correctly for these individuals *without using non-census sources*. Part of the reason for this can be attributed to the lack of clear instructions to enumerators about the collection of occupational information, with the exception of the case of railroad clerks (Healey, 2011; U.S. Census Office, 1880). Unfortunately, instead of recognizing the problem both in the data set and in the documentation for the industry variable, the NAPP project has made incorrect industry imputations for hundreds of thousands of workers in generic occupations.

A brief analysis of the occupational/industry code combinations in the entire 1880 data set makes clear how serious a problem this is. Taking the case of blacksmiths, 99% of the 177,193 individuals are given an industry code of 817 for “miscellaneous repair services.” Only 373 blacksmiths are coded as railroad employees, 142 to different iron and steel related codes, and a mere five to coal mining. Such figures are entirely incorrect and extremely misleading. For example, a single anthracite mine (the Diamond) out of more than 100 in one of the four anthracite coalfields, employed four blacksmiths and three blacksmith’s helpers in mid-1880 (Diamond Payroll, 1880), so across both the anthracite and bituminous mining sectors, thousands of blacksmiths would have been employed and the same would have applied to the other sectors named above. For machinists, the situation is similar. Of 97,424 workers in this occupation, 97.3% were

coded to the mysterious category of “miscellaneous machinery” (code 358) and only 1,177 or 1.2% to the railroad sector. However, the 1880 census special report has a specific breakdown of numbers of machinists (unlike blacksmiths), and it states that 22,766 of them worked for the railroads across the United States (U.S. Census Office, 1883b). This means that more than 21,000 of this sub-group are misclassified. Likewise, the Mine Inspectors’ reports for 1880 record 813 “outside mechanics” employed at mines in the northern anthracite field (calculated from data in Inspectors of Mines, 1880). While this definition may not exactly equate to machinists, according to the NAPP data set there were no machinists at all who worked in coal mining.

Prior to 1880, the enumerators’ instructions were no more specific than in 1880, despite a minor caution against using “machinist” if a more precise description could be given (U.S. Census Office, 1870, 14). After 1880, the instructions for 1890 and 1900, which are largely the same in both years, suggested in places the need for accurate qualification of job titles, by means of isolated examples such as “railroad laborer” or “carriage blacksmith,” but there is no clear recognition of the systematic need to distinguish generic from industry-specific occupations. This can be seen in the complete omission of generic trades (as opposed to laborers) from the list of steam railroad occupations (U.S. Census Office, 1900, pp. 32, 36). As a result, the same problems with underreporting are to be expected both in the earlier and later censuses. This is unequivocally demonstrable in the case of 1890, as there was also a special report on transportation in this year (though not in 1900). The special report gives a figure of 750,017 employees, but the Edwards Report only counts 462,213 (Edwards, 1943, p. 109; U.S. Census Office, 1895, p. 130). This special report figure for 1890 is also greatly in excess of Edward’s 1900 figure of 582,150, at a time when railroad employment was still expanding, so the underreporting issue was still not resolved by this date. Although attention has been focused on generic skilled tradesmen in the previous examples, unskilled general laborers also make a substantial contribution to the overall problem, because of their comparatively large numbers, and the likelihood that their industrial sector was also not recorded by the enumerators. Further to this, laborers in irregular employment, say in railroad construction, may well not have identified themselves as part of the railroad industry, even if specifically questioned to that effect.

The much wider implication of these problems is that the NAPP/IPUMS 1% sample census data sets for other census years after 1850 (excluding 1890) are also subject to the same difficulties of interpretation and inaccurate assignment of sample individuals to industrial sectors. Any research findings based on these specific industrial codes may therefore be very much in error, and these errors are unlikely to be consistent between different industrial sectors. The potential impact on studies of inter-sectoral mobility is substantial. The same applies to inter-censal analyses of changing occupational/

industrial structure based on aggregate statistics, or of detailed occupational mobility, based on linked samples derived from NAPP or IPUMS data sets. This is apparent, because there is no means of determining from census records alone, whether occupational information about given linked individuals was recorded in the same way in successive censuses, so workers may appear to be railroad employees in one census but not the next, when their employment status did not actually change. Further to this, sampling from undifferentiated occupational groupings, when those same groupings actually contain different sub-populations of individuals in different industrial sectors, may be a source of concealed bias in statistical studies. For example, it has already been shown that railroad machinists and blacksmiths in Baltimore in 1860 had different socio-demographic characteristics than their non-railroad counterparts (Healey, Thomas, & Lahman, 2013). It is therefore most important that these coding issues are more widely discussed and analyzed, to prevent inappropriate analyses being undertaken that generate misleading or false results. A further inference is that historical census data sets, standardly viewed as “givens” for secondary data analysis, should more accurately be viewed as “works in progress,” resources whose data quality needs to be enhanced progressively over time, by means of comparison with other sources, to increase the confidence that can be placed in analytical results derived from them. This is not a welcome finding for research funding bodies, who would doubtless have wished that researchers could capitalize on their past investments in large data sets without the need for ongoing expenditure on quality improvement. It also leaves some individual researchers in a quandary, as it is now clear that the coded data presently available cannot support certain types of analyses that would previously have been deemed viable. They can either restrict the scope of their work (e.g., by avoiding use of industrial sector codes) or shoulder the rigorous additional burden of making the required data quality enhancements using non-census sources. While the latter may be a feasible strategy for well-resourced work with limited geographical coverage, it is infeasible for individuals wishing to engage in larger-scale studies. Also, in the absence of any agreed approach to the use of non-census sources or how any re-coding might be undertaken, there is serious risk of incompatibilities quickly arising between studies, which will greatly hinder future comparative work. Where studies only make use of very broad occupational categories (e.g., Ferrie, 2005), the impact of these detailed problems may be lessened, but it can no longer be assumed that they do not exist.

## Requirements for a New Coding System

To address this unwelcome situation in a systematic manner, a new approach is required to the problem of quality enhancement of existing historical census data sets, such as the NAPP 1880 census. This involves several initial steps. The first of



these is to provide an overview of the main types of non-census sources that may eventually contribute to the re-coding process. The second is to evaluate what new developments, in terms of coding capabilities, are required to mesh together census and non-census sources. The third is to identify a suitable computational methodology or methodologies that will support these new capabilities. The fourth is to identify operational considerations that could facilitate the take-up of new coding system capabilities, and finally, there is the need to outline future possibilities for systematic re-coding projects (e.g., of specific industrial sectors) of sufficient substance to demonstrate unambiguously the full nature and extent of the data quality problems for the sectors in question, and to provide guides to assist subsequent projects aimed at other sectors. The main emphasis of the present discussion will be on the first three of these steps, followed by a brief commentary on the remaining two stages, the implementation of which lies, at least in part, in the future.

The first question to address is which other non-census sources are available to assist with census (re-)coding. A range of these can be identified in the U.S. context, but they vary widely in their temporal, geographical, and sectoral coverage and indeed their degree of comprehensiveness, even for specific locations and time points. Among the most obvious candidates are city directories, company payrolls, marriage and death records, and naturalization records. Less obvious candidates would include the harrowing industrial accident records found in state railroad commission reports and mine inspectors' reports. While space precludes a detailed survey of these sources, several brief comments serve to highlight relevant issues. The census has the enormous advantage of relative geographical comprehensiveness over a broadly comparable time interval (the concept of a precise census date was not well-developed in earlier years), and provides information on age, family, and household status, occupation and birthplace. Marriage and death records will provide a subset of this information possibly with links to parental names. Company records, such as payrolls, being employment-focused, lack much of this information, including age-related data (though this may be found in employee card indexes). However, this is offset by the detailed work history information they contain. Directories, though largely confined to urban areas, have varying degrees of comprehensiveness for the populations they served, lack age or family data, but provide addresses and often contain valuable employment-related information for multiple time-slices falling between census years. The potential research benefits of being able to combine data about individuals over time and space from these and other relevant sources are easy to see, though the practical problems of achieving the required data linkage in a reliable manner may be quite another matter.

To examine some of these sources in more detail, experience with city directories, for example, suggests they are most informative for occupational purposes in the 1850s to

1870s, rather than in later years, and the larger the city the less informative, owing to pressure on space in individual volumes. By more informative, is meant more likely to provide not only an occupation for each individual, but also an industrial sector or even specific manufacturing establishment/company department (e.g., foreman of the car repair shop of a specific named railroad). Comprehensiveness of population coverage probably increased over time, as directory compilers became more organized and better funded, though systematic studies of this are largely lacking (Goldstein, 1954). Payrolls are usually much more detailed, though far more sporadic in space and time. Thus, only a small fraction of 19th century anthracite mines have surviving payrolls, and regrettably even fewer railroads, but the documents that do survive, will reveal much finer job subdivisions than "coal miner" or "railroad hand." They may also indicate the department of the company in which employees worked, and provide information on how they were paid (piece-work or hourly) and the regularity of work over shorter or longer periods, depending on the length of surviving records. The clear advantage of payroll records, and indeed industrial accident records, because the information is firm specific, is that they are guaranteed to address the problem of identifying generic workers in specific industrial sectors at particular dates. This is not standardly the case for city directories, although some early volumes do contain a good deal of the requisite information.

Setting aside questions about relative ease of processing of printed versus manuscript sources, and the major topic of nominal record linkage between different sources (for a review, see Winkler, 2006), which are beyond the present scope, key requirements for an occupational coding system that facilitates re-coding of census records using non-census data can now be identified, based on the range of information that may be available in different types of non-census sources. First, and most importantly, the system must enable workers in generic occupations to be "tagged" with their specific industrial sector, where known. Second, it should extend beyond the rather general occupational categories favored by census enumerators and "genealogical" sources, such as marriage records, to encapsulate the greater range of employment information provided by payrolls, industrial accident records, and many early directories. This information includes detailed job titles, major and minor sub-divisions within companies and whether employees were engaged in construction work or activity related to production/operation. Thus, the system should be able to distinguish blacksmiths involved in railroad construction from those employed in the operation of rolling mills in the iron industry. Such distinctions are impossible to make with the version of the HISCO coding system used for the U.S. 1880 census and the samples from earlier and later censuses. This is the main reason why a new system is required. Further to this, however, if non-census sources are to be used, it is sensible to abstract as much relevant information on employment structure from

them in a single pass as possible, to avoid the need to keep referring back to them for more detailed information. In this sense, the resulting occupational/industrial code then serves as a kind of employment structure index to the archival source (e.g., a payroll), in addition to its main function as a classificatory device. This proves to have wider implications, as will be seen below. At the same time, the wide usage of the HISCO system means that backward compatibility with it should also be provided by the new system. Relating detailed job titles to the more general HISCO categories also obviates the necessity for a separate look-up table of individual titles. Finally, unlike some of the older systems developed in the pre-Internet era, it will be assumed that the system can take full advantage of a range of readily available digital and database-related technologies, including Web connectivity.

### Focus of the New Coding System

As Wrigley has correctly observed, there is no right or wrong in terms of coding systems, but each will have a particular class of problems to which it is especially well suited (Wrigley, n.d.). In his case, Wrigley adopted a focus on the distinction between primary, secondary, and tertiary (PST) sectors in the economy, because of a particular interest in the changing relative importance of these sectors over the long term as the Industrial Revolution progressed. Herschberg, in contrast, seems to have sought an all-encompassing census coding system, with a certain focus on industrial sectors, though his precise aims are not very clearly articulated, and there is no recognition of the problems caused by lack of specificity in the occupational/industrial sector enumeration of generic workers (Herschberg, 1976). In the present system, a particular, though not exclusive, focus is on the question of occupational and geographical mobility of heavy industrial workers. This has many facets, as workers can change jobs within and between industrial concerns in the same sector or utilize existing generic skills in new ways by changing industrial sector, for example, a machinist moving from the mining to the railroad industry. Such occupational movements, which may be upward, horizontal, or occasionally downwards, in terms of the job and remuneration hierarchy, may or may not be accompanied by geographical mobility. Changes of location in pursuit of career advancement have been characterized as a major feature of the "American Dream" (cf. Ferrie, 1995, 1999) and thus the problem of mobility touches on a wide range of debates about immigration and 19th century economic growth (Thomas, 1973).

A focus on occupational mobility has also been the preferred approach in relation to the U.S. Department of Labor Dictionary of Occupational Titles (Miller, Trieman, Cain, & Roos, 1980). A substantial body of work on coding systems, both for the U.S. census and for inter-agency work within the U.S. Government, undertaken in the first half of the 20th

century and summarized by Palmer (1939), concluded that rigorous and consistent classification based on a distinction between skilled and unskilled work was infeasible. In contrast, both Herschberg (1976) and Morris (1990), in the U.K. context, have stressed the need for any coding system to reflect both occupational characteristics and industrial sector affiliation. This is particularly significant, as the NAPP-modified HISCO system (Roberts et al., 2003) does not do this as part of its structure, though some specialized occupations will tend to be associated with particular industrial categories.

### General Design Criteria for the New System

There are four general design criteria for the system that need to be explained prior to detailed treatment of the individual components of the overall structure. The first derives from the requirement stated above that the system must be able to code both census/vital registration records and occupational data from industrial/company records. The HISCO system was *only* developed on the basis of the former type of data (Van Leeuwen, Maas, & Miles, 2004). Second, as the coding system is intended to capture the more detailed information derived from company records and some city directories, it should be able to do this in a way that facilitates analysis of occupational mobility within industrial sectors, as well as among them. None of the systems in current use support this type of analysis. The third criterion is that a single code will be used, rather than a multi-component code, as in the case of the PST and Herschberg systems. Likewise, as noted earlier, NAPP data sets have separate industrial and occupational codes. The present codes therefore span both industrial sector and occupational categories. While this may initially appear to hairsplitting, it proves to be of major importance in implementation terms. The final criterion is that the coding system is strictly hierarchical, even if this makes it appear "verbose" in places. The combination of these two criteria allows this system, unlike its predecessors, to be used in the exacting implementation environment of a properly constructed data warehouse, where in-built Online Analytical Processing (OLAP) functions allow automated aggregation and dis-aggregation of categories and sub-categories of data, based on different hierarchical levels in the occupational coding of individual records (Healey, 2011). Within the overall structure, there are eight hierarchical levels, as indicated in Table 1. As with all such systems, there is no intrinsic validity to this number of levels, but it was found in pilot tests on different industry sectors that it allowed for considerable detail about the nature of employment categories within industrial concerns to be captured, as well as inter-sector differences. This is obviously of particular value when working with company records, which have not formed the basis of the better-known systems reported in the literature. Each level is represented by a specific number of digits

**Table 1.** Hierarchy of Levels in the New Coding System.

Level	Description
8	Major industry categories
7	Major industry sub-categories
6	Production/construction breakdown
5	Company sub-division or sub-type breakdown
4	Company detailed operating/production sub-division
3	Main job type divisions (incorporating highest level of NAPP-modified HISCO codes)
2	Job sub-divisions (incorporating middle level of NAPP-modified HISCO codes)
1	Detailed job titles (incorporating full individual NAPP-modified HISCO codes, where available)

Note. NAPP = North Atlantic Population Project; HISCO = Historical International Standard Classification of Occupations.

in the code, the number of digits varying between levels, as required. In the first version of the system, the overall code is 14 digits long. Simple database string manipulation functions allow any level sub-code or combination of sub-codes to be selected, in addition to the entire code, so it is very flexible in use. The individual levels will now be examined in more detail, and the full implementation can be examined online (Healey, 2010).

### Individual-Level Sub-Codes

From Table 1, it is apparent that the numbering of levels runs from the highest at 8 (the most general) to 1 (the lowest and most disaggregated). *Level 8* (2 digits) is the broadest level of industrial classification in the system. Comparison with other systems shows a great variety of methods of classifying industrial sectors. For 19th century purposes, Herschberg (1976), for example, found earlier releases of the present day Standard Industrial Classification (SIC; U.S. Securities and Exchange Commission, 2011) to contain anachronisms. The SIC also omits categories now unimportant, but previously extremely large, for example, anthracite coal mining. Equally, the industrial codes used by NAPP correspond to the system set out by the Bureau of the Census in 1950 (U.S. Bureau of the Census, 1950). Necessarily designed primarily for 20th rather than 19th century census coding, this system is subject to exactly the same criticisms leveled at the SIC codes above, for example, it merges anthracite and bituminous mining, fails to distinguish canal transportation from other kinds of transportation on inland waterways and groups blast furnaces, steel works, and rolling mills together into a single code. It also provides guidance on which occupations fall within which industrial sectors, but in the present context some of its specific directives are very problematic, such as the instruction to include mine laborers under the very general and uninformative category of “operatives and kindred workers,” rather than under the still too broad category of “coal mining” (U.S.

Bureau of the Census, 1950, p. xx). Given the rather limited attention, apparently given in this system to proper hierarchical structuring of the codes into major and minor categories, and the problems listed above, there is little to be gained (and a good deal to lose) by trying to match the 1950 industrial codes to those used here. Indeed, by deliberately making the present codes quite different, it reinforces the argument above that the 1950 codes, as deployed by NAPP, contain limited useful information which reduces their value for detailed analytical purposes. As the coding system has primarily been established for heavy industrial workers, in the first instance, the Level 8 categories currently implemented include coal mining, iron and steel, and transportation. These categories are more than sufficient to demonstrate the structure and operation of the overall system.

*Level 7* (1 digit) provides appropriate sub-categories of the main industries. Thus, coal mining is currently divided into anthracite and bituminous mining. To the non-specialist, this may appear unnecessary. However, the marked difference in the ratios of “inside” (underground) to “outside” workers and the major differences in the range of outside occupations between the two sub-categories provide the rationale for this. Anthracite coal requires extensive processing and sorting into size fractions in a “coal breaker” after being hoisted to the surface, whereas bituminous coal does not, so the range of occupational types and the number of outside employees is much larger in the former case (DiCiccio, 1996). This level also distinguishes primary iron manufacture from steelmaking, a necessary distinction even after more integrated iron and steel works began to appear from the late 1860s to early 1870s onward in the United States (Temin, 1964). In the transportation category, railroads are currently implemented, but additional sub-categories will obviously be required in future, for canal workers and employees in other forms of navigation on inland waterways, both major rivers and the Great Lakes.

*Level 6* (1 digit) makes the important distinction between workers who are engaged in productive enterprise and those who are responsible for the construction of mines, mills, and conduits of transportation, such as railroad lines or canals. There are many reasons for wanting to maintain such a distinction. For example, construction activity, following on from investment decisions, was broadly cyclic in nature and could attract hundreds or even thousands of workers to specific localities for extended periods of time, running into years for large mining or railroad developments (Healey et al., 2013). However, once construction was completed and productive operations began, the nature of employment opportunities could change rapidly, necessitating substantial geographical mobility of labor, in search of continuing work. For example, various stages of wooden coal breaker construction required large numbers of carpenters, but once a specific mine was in operation, the demand for these generic tradesmen at that location dropped sharply (and standard carpentry skills were not immediately transferable to

the specialized job of “inside timber-man,” responsible for shoring up underground mine workings). Accurate identification of construction workers is much more likely to be possible when payroll information is available, although justifiable inferences can also be made under appropriate circumstances, for example, when a city directory lists a worker employed at a specific mine and it is known from other sources that this mine is under construction at the time in question. Examination of the NAPP 1880 data set suggests that industrial construction activity is very largely hidden from view, not least because the vast majority of the large cadre of construction laborers are simply reported under the generic laborer codes.

*Level 5 and Level 4* (each 1 digit) enable increasingly detailed tracking of the main and detailed sub-divisions/department within a company where individual employees worked and, depending on the particular industry sub-sector, their relationship to their employer in contractual terms. Taken together, they represent an important contribution to coding methodology, because such information is largely lacking from all the standardly used systems (even Herschberg’s extensive coding structure does not approach the level of detail captured here). Where the necessary data to inform the coding is available, these levels can be used to analyze horizontal and vertical occupational mobility of workers within a single large firm. This mobility may also have an important geographical component if the firm operates on multiple sites. While the best choice of level at which to make certain sub-divisions may be arguable, pilot testing indicated that *Level 5* should be used for major sub-divisions of productive activity and *Level 4* for more detailed sub-divisions. Not all industry sectors would necessarily require use of both levels, in which case a “pass-through” mechanism is deployed, whereby the classification from the level above is simply repeated one level down. This is an example of how the code may appear verbose or slightly redundant in some instances, but the advantages of a standardized hierarchical structure, common to all industries, greatly outweighs any minor lack of elegance in implementation. Thus, at Level 5, the iron industry sub-divides into different types of activity, which may or may not be found on the same site. These include blast furnaces, rolling mills, and foundries. At Level 4 in rolling mills, for example, contract or piece-workers are separated from workers paid a daily wage. A somewhat analogous process applies to anthracite mining, where Level 5 separates inside from outside workers, and with Level 4 inside workers, “company men” paid a daily or monthly wage are coded differently to contract miners, paid by the car of coal mined and loaded. These distinctions are immediately apparent when processing anthracite mine payrolls and present very few problems of interpretation under these circumstances. In the absence of such archival sources, it may be very difficult to benefit fully from these coding distinctions, as neither the census nor city directories report at this level of detail and the undifferentiated term of “coal miner”

or even just “miner” is normally found. For railroads, Level 5 currently serves as a pass-through, but Level 4 is very important because it is used to reflect the standard departmental and accounting breakdown that came to be used on most railroads during the 19th century. The main sub-divisions are Conducting Transportation, Maintenance of Way, and Motive Power. The first of these is concerned with the daily operation of train services, both freight and passenger. Maintenance of Way covers all track and depot repairs and Motive Power refers to the building and repair of locomotives and rolling stock and all associated engine-house and machine shop functions. An early reference to such a classificatory scheme, crediting the Georgia Railroad as the originator, can be found in the 1850 annual report of the Pennsylvania Railroad (Pennsylvania Railroad, 1851). Initially, “maintenance of cars” was treated as a separate department, but over time this was merged into the “motive power” heading. With minor variations, the scheme was widely adopted by the 1880s. For coding purposes, the scheme has been extended to include a category of “General Administration,” as, by convention, these administrative employees and company officers were not included in the threefold sub-division above. There are also “catch-all” categories for employees for whom detailed information is not available. Interestingly, in cases where payrolls are available for years prior to the adoption of the scheme by a given railroad, aggregation of individual lower-level job codes (see below) to these Level 4 categories has already been shown to provide a very effective standardized means of analyzing the changing employment structure of a given railroad over time. By extension, it can also be used effectively for systematic comparison of the structure of different railroads (Healey et al., 2013).

The lowest three levels (1-3) are designed to provide compatibility with the NAPP-modified HISCO coding system, while also enabling this system to be extended substantially to include the much wider range of specific occupations found in non-census records. The maximum possible compatibility is sought, subject to the constraints of a strictly hierarchical coding system. While HISCO very largely meets these requirements, there are occasions where lack of consistency or other considerations required slight modifications to be made or minor re-naming to take place. The effects of these changes are limited, however, so the overwhelming majority of HISCO codes can be readily identified within the new system. However, some modifications that do require highlighting at the outset are that the codes as they appear on the NAPP list have an additional digit to the left and two to the right of the code in the present system. Partly, as a consequence of this, the division of the code into levels for hierarchical decomposition differs slightly from that in the original HISCO scheme. In the latter, codes have the form 9.99.99, where 9 is the placeholder. The highest level (one digit) represents major groups, the middle-level (two digits) minor groups, and the lowest level (two digits) the individual



occupational categories (Van Leeuwen et al., 2004). In the present scheme, the structure is 09.9.99900, using zeroes to show the new digits, but, as can be seen, the five HISCO digits remain, though divided slightly differently.

Examining the new expanded structure above, *Level 3* (2 digits) corresponds to the HISCO major groups, which can be given broad labels such as “Professional and Technical” (sub-code 01) or “Production, Construction and Transport” (sub-code 09). The additional digit is used to provide a 10th category, not found in HISCO, which cover the varied instances where occupation is lacking, but useful information is still provided, for example, for retired persons. This permits inclusion within the code of data that in the HISCO scheme is separated out into multiple additional variables such as STATUS or RELATION, that do not form part of the main code structure (Van Leeuwen et al., 2004). *Level 2* (1 digit) provides a middle-level breakdown, corresponding to the first part of the HISCO minor grouping, although it is argued that the categories used here are much more obvious than in the NAPP/HISCO code list (NAPP Code List, 2013) and have been given clear labels. For example, under the Level 3 sub-code 09 for inside company workers in anthracite mines, the Level 2 sub-categories are mine development workers (095), stationary engine operators (096), transport equipment operators (098), and workers not elsewhere classified (099). Comparison with the NAPP code list immediately shows how the present system aims to preserve compatibility with HISCO wherever possible, but does not allow the latter to restrict necessary extensions/modifications to the code structure. Thus, HISCO has no equivalent of mine development workers, as it lacks sufficient detail to code them separately. Consequently, the codes in HISCO beginning with 95, which relate to various types of building work, are deployed *in this specific branch of the system concerned with mining*, to refer to “building” work inside the mine, which is necessary for mine development underground, but which is not actual mining of coal. In contrast, major types of stationary engine operators can be readily identified in the NAPP list as codes starting with 96, so these can be deployed directly in the current system, but with the additional interpretation at this point, derived from the higher levels of the code, that they are stationary engine operators working inside a mine. The same type of argument can be applied to the other Level 2 sub-categories in this example. At *Level 1* (5 digits), considerable additional detail about specific individual occupations can be coded. The first three of these digits comprise the remainder of the full HISCO code for the occupation in question, assuming it appears in the NAPP code list. The final two digits provide the ability to add extensive refinements of the HISCO occupational categories, based on additional information from company records, or to add new codes for jobs that are not found in the NAPP code list. Thus, to continue with the stationary engine operators example for inside mining employees, the full NAPP/HISCO code for “Stationary engineers and engine

men” is 96230. This appears as an entry in the present system, as 09623000, to show that the full NAPP/HISCO code is embedded in this more expanded code. However, other sub-categories, such as “donkey runner” (for operator of auxiliary “donkey” engine) or pump-man, neither of which have NAPP/HISCO equivalents, appear as sub-codes of stationary engine operator, that is, sub-codes 09623010 and 09623040, respectively. For clarity, their full codes, showing all levels, are 10111109623010 and 10111109623040. As all digit positions are fixed in the code, it is a trivial database operation to select all codes that refer to stationary engineers, wherever they appear under different branches of the system, such as inside mines, in blast furnaces, or in railroad motive power divisions. This means that the system can immediately be used to generate aggregate occupational statistics that are compatible with those that can be derived from the NAPP census data sets. Table 2 provides a further illustration of the overall structure of the coding system, using an example of workers in a rolling mill and following this specific occupational sub-tree down through the eight levels. The three dots “. . .” in several cells in the table indicate that there are additional categories at the level in question, which are not shown for reasons of clarity.

### Expanding the Range of Occupational Titles—Coding Issues and Data Sources

Within the coding system, the occupational description that accompanies each code at the lowest level may be derived from a variety of sources. The more generic job titles correspond to those in the NAPP/HISCO system and are readily identifiable as a result. However, as previous studies have pointed out, the NAPP/HISCO coverage of occupational types found within different industrial sectors is very variable. Thus, there are 18 different NAPP/HISCO codes specifically associated with railroads, but only two codes for the mining industry (71120 “miners,” which unhelpfully includes both coal and ore miners, and the general and very rarely used 71190 “others working in mines and quarries”). The rather unclear code 71200 “mineral or stone treaters” might also be partially relevant, but it is not obvious from the code list whether this relates to mining activity as normally understood (see below). This comprehensive failure to recognize the occupational complexity of the mining industry was particularly striking to the present writer, because earlier unrelated analysis of anthracite mine payrolls in the 1880s and 1890s had already shown that about 200 distinct types of work could be identified without any difficulty. More recent work on payrolls from the 1860s, as this coding system was developed, and mine inspectors’ reports from the 1870s onward has further extended this list of job types to around 300 in total, not all of which are yet incorporated in the system (Healey, 2013).

Also, in an attempt to clarify the use of the code 71200 in the NAPP 1880 census data set, the detailed occupational

**Table 2.** Example Occupational Sub-Tree Leading to Detailed Categories of Rolling Mill Workers.

Level 8	Level 7	Level 6	Level 5	Level 4	Level 3	Level 2	Level 1
Coal mining							
Transportation							
Iron and steel	Steel						
...	Iron	Construction					
...		Production	Blast furnace	Mill day workers			
			Rolling mill	Mill contract workers	Mining/Metal manufacture	Metal processors	Heater rail mill
			...	...			Puddler
							Roller bar mill
							Rougher guide mill
							...

transcriptions for workers with this code were examined on a sample basis. This revealed interesting and important lessons in the present context, both for those engaged in the coding of census occupations, and subsequent users of the coded data. As noted earlier, a unique characteristic of anthracite mines was the requirement to prepare the coal for market in large “breakers” on the surface (Hudson Coal Company, 1932). This meant that large numbers of boys (and some elderly ex-miners) were employed at each anthracite mine to help with this preparation process, which was only partially mechanized in the 19th century. Many thousands of these “breaker boys” or “slate pickers” (so-called because they removed rock or “slate” from the coal before it was loaded into railroad cars) were employed in the anthracite coalfields, but not in the bituminous mining regions, where breakers were not required. In the NAPP data set, these boys have almost all been coded to 71200. This separates them from “miners” per se and their NAPP/HISCO occupational description tends to mislead rather than inform, as they are never described as “mineral treaters” in the mining literature. The rationale for this coding decision is also much more apparent following the above explanation of the work of breaker boys, than it is when examining a code list to identify potential mining industry employees, as one group among many chosen for analysis. Put another way, if it is necessary to utilize detailed industry knowledge and the original occupational transcriptions to understand the use of a code, then it is probably not a very effective numerical shorthand. Further to this, from an industrial sector perspective, use of 71200 confounds coal mining-related activity with unrelated stone-dressing in quarries or mineral ore processing to an unknown degree in any county data set, though it is fortunate that anthracite is only mined in a very limited number of U.S. counties. In contrast, these breaker boys are always reported as an integral part of the anthracite mining industry, both in mine inspectors’ reports and in company payrolls themselves. The present system therefore identifies them under the code 10112209900220, which specifies that they are anthracite industry employees paid by the day to work

outside the mine in the coal breaker. To avoid perpetuating the confusion generated by the 71200 code, a different code not used by NAPP/HISCO, namely 99000, has been used here, to provide the basis for sub-codes to match the lengthy list of job types found in coal breakers.

Although the NAPP/HISCO codes are more informative and differentiated for the railroad than the mining sector, the 18 codes still only represent a small fraction of the job types actually found in railroad employment. They are excellent for trainmen, who would fall under the “conducting transportation” heading, but not for the generic trades more prominently found in railroad shops, who were classified under “motive power.” To remedy this deficiency and provide a more balanced coverage of occupational types across the sector, two main sources were used to provide a list of about 300 job titles. These include the published payroll lists of the Baltimore and Ohio Railroad, which cover the years 1842 to 1857 (Baltimore and Ohio Railroad Company, 1842, 1852, 1858) and the reports of the Pennsylvania Bureau of Industrial Statistics (PBIS; Secretary of Internal Affairs, 1877, 1881). The latter are especially important, as they contain details of the occupational structure, including job titles, of large numbers of different railroads, large and small, within the state during the 1870s and 1880s. Although no set of listings can be considered exhaustive, comparison of the returns for the different railroads in Pennsylvania with those of the Baltimore and Ohio in Maryland and Virginia provides an excellent basis for the railroad job titles to be coded in the present system. Most importantly, unlike NAPP/HISCO, this list is not biased toward the trainmen, and provides good coverage not only across all the three main departments of railroad operation but also extending to categories of railroad construction workers, as these were recorded in the large payroll list of the Baltimore and Ohio in 1857.

The PBIS returns are not limited to railroad reports, and they also provide details of employment in bituminous coal mines, in primary iron and steel manufacturing concerns, and in rolling mills. These have been utilized within the coding system, and, as would be expected, a distinction is made between

workers with otherwise similar titles, depending on whether they were employed at blast furnaces or in foundries and so on.

Overall, across the various sectors, the system currently has 2,372 occupational entries, each of which has the eight levels of the hierarchical code structure attached, making nearly 19,000 code values, though the number of different job titles is much smaller (814), because generic occupations, such as machinist, will be found under several sector and sub-sector headings. Any user would only use the 2,372 Level 1 codes; the others are used by the data warehouse in which the code system is embedded. In any primary data source, a number of individuals will have incomplete employment attribution, for example some blacksmiths may be known to be employed by a given iron and steel works, but it is not stated whether they worked at the blast furnace or the foundry. To allow for this, every level of breakdown enables workers without a more precise lower-level classification to be coded as “unspecified” and the numbering convention is standardized for this (codes end in “95”). Examples would be “Inside Mine Worker Unspecified Occupation” or “Bloomery Worker Unspecified Occupation.” Further to this, there are a group of codes for generic occupations in industries other than those currently handled in detail. These codes still convey slightly more information than their HISCO equivalents, as identified workers in these occupations in the main heavy industry categories have already been separated out. For example, they allow coding of individuals, who are recorded in city directories as carpenters in furniture manufacturing plants. This flags the existence of some additional information in the original sources about these workers, should that be needed. If such industries as furniture are specifically coded in future, these workers can be retrieved from the data set and given more precise codes at that time. Where no industrial sector attribution is available, separate codes again are available to cover individuals who are simply recorded as “carpenters” or “blacksmiths,” and there are the usual “catch-all” codes for individuals who lack the information to enable them to be otherwise usefully classified.

While the hierarchical structure closely guides the coding process, the opportunity has been taken to standardize certain code components to facilitate the retrieval of individuals with particular employment characteristics that may span multiple occupations. For example, all titles that include the word “helper” or “assistant” end in the digit “1” at the lowest level. However, as an assistant master of machinery is a very different level of job than a blacksmith’s helper, the fixed number of digits in each code means that the simple application of a format mask to the code allows assistants in supervisory posts to be distinguished easily from helpers in standard trades. A similar convention applies to apprentices, all the codes of which end with the digit “2.” This approach equates to that used by Herschberg, but differs from HISCO, where such qualifying information would have to be found in the separate STATUS code.

## Operational Use of the New Coding System

As noted at the outset, the new coding system is a key component of a future process of what might be termed *occupational accuracy improvement* for historical census data sets. This is a long-term goal, which must be approached systematically, if it is to have any likelihood of success. It will undoubtedly require work by a number of research groups and will increasingly deploy the methods of citizen science or crowd sourcing, whereby large numbers of individuals contribute limited packages of work that amount in total to a major research contribution. However, crowd sourcing approaches take time to establish and gather momentum. In the interim, a pilot project has been launched by the present author, focused on major urban centers in the American Manufacturing Belt, such as the cities of Cleveland, Ohio and Scranton and Pittsburgh in Pennsylvania to examine the potential for extracting additional information from city directories, and where available, company records, and matching it to census records. As would be expected from the foregoing, the emphasis is on the railroad, mining and iron and steel sectors, which, to differing degrees, formed a major part of the industrial base of these and similar cities. This combination of specific geographical locations and selected industrial sectors means that the resulting data sets will be valuable in their own right as case study examples and become increasingly so, as the scope for comparability between cities and sectors grows over time. Further to this, as employment in these sectors was heavily concentrated in and around these cities in the Manufacturing Belt as late as 1880 and beyond, these new data sets will progressively create growing collections of re-coded data. One of the future uses of these collections will be to determine the possibility of applying correction factors to other data still to be re-coded, to improve estimated findings based originally on the latter. When re-coded occupational/industrial data are added in bulk to an existing census data warehouse, it will also become possible to compare tabulations based on the original codings with those based on the quality enhanced, re-coded data.

Preliminary findings indicate the feasibility of abstracting many thousands of new data records from city directories, though accurate matching to census records is a resource-intensive process. It has also been found, as was anticipated, that managing lengthy numerical codes in the course of manual coding of directory and other data, is problematic for the personnel involved, though the subsequent computational use of the numerical codes is very straightforward and effective. To bridge this operational gap, while retaining the considerable database/data warehouse benefits of the numerical codes, a structured set of mnemonic character abbreviations has been devised, and the railroad sector is being used as a first test of their effectiveness. These abbreviations correspond to the relevant sections of numerical codes,

so automated conversion can be undertaken, but unlike the latter, early experience shows they are finding ready acceptance for manual coding purposes. This is facilitated by the fact that a limited group of occupations, such as brakeman (mnemonic = rrb) and conductor (mnemonic = rcn) in the railroad sector, account for a significant proportion of all employees, so the most frequently used mnemonic codes can be memorized quickly through repetition. After testing is completed, these mnemonics and the conversion tables will be made publicly available in the same way as has already been undertaken for the numerical codes, so other research groups can utilize them if desired.

## Conclusion

The system resolves the key shortcomings of the HISCO coding system, by encompassing it within a much more sophisticated structure that allows comprehensive coding of data from both census and non-census sources, to a level of detail compatible with that provided in the original source documents. As it includes the HISCO codes, it maintains a very high level of compatibility with that approach, yet avoids the necessity for separate look-up tables, as provided by Wrigley for the PST system (Wrigley, n.d.). However, as Wrigley has helpfully provided such tables, this also means that a high degree of compatibility exists with that system also, by deploying these intermediate tables in conjunction with standard database queries. The new system has the important ability to standardize employment data from company payrolls and other industrial archives, as well as coding census and vital registration records. The fine breakdown of employment characteristics that it provides offers a much more nuanced approach to the analysis of inter-departmental, inter-sectoral, and geographical mobility than is possible using other coding systems. A further important motivation for its original development was to allow comparison of the demographic characteristics of sub-populations of generic workers in different industrial sectors and this capability has already been demonstrated in a small case study in Baltimore (Healey et al., 2013). Further work on re-coding selected data from the 1880 census in a data warehouse context is planned to develop this approach on a larger scale.

## Future Development of the Coding System

While an exhaustive list of occupational titles for different U.S. industrial sectors in the 19th century is probably an unattainable goal, a very comprehensive list can eventually be arrived at through comparison of multiple sources, both printed and archival. While much of the groundwork for this has been laid for the heavy industrial sectors, more can still be achieved by incorporating data from two late 19th century sources. The first of these is the report of the Commissioner of Labor (1890), on railroad labor, which contains complete

lists of job titles for a small sample of major railroad systems across the country. While the vast majority of the common titles listed in this source are already in the system, some of the less common ones are not. The second is the Weeks Report and the associated database (Meyer, 2004; Weeks, 1884), which contains a large number of job titles in different industrial sectors, though it makes no claim of completeness. Ideally, more payroll information would also be incorporated, although payrolls both for railroads and large iron/steel works are surprisingly difficult to locate in any quantity for the latter part of the 19th century (see Knowles (2013) for examples of the use of earlier iron company records).

Another issue is that, over time, certain titles fell into disuse, or persisted in some regions but not others, or the nature of the work activity that they represented changed quite significantly, as technology moved forward. The original largely European focus of the HISCO system raises further questions, as many U.S. occupational titles differed from their British equivalents (measured in terms of the tasks involved) or the English translation of French or German terms may not correspond to U.S. usage of the word in question. Hence, there is a wider research agenda, not well-articulated in the literature, than the narrower field of comparability between occupations recorded in censuses internationally. As the new coding system aims to include original job titles, not a subset of standardized categories, it also has the potential to act as an index to more extensive textual resources that describe the “task lists” of apparently equivalent jobs in different companies, sectors, and locations and how these evolve or change over time. This would be a useful extension to the helpful occupational descriptions and illustrations already provided online for HISCO codes (History of Work Information System, 2013). Some progress in this direction has already been made, in terms of identifying published descriptions of railroad jobs in a range of companies from the 1860s onward. While time-consuming to develop such a textual resource, it is straightforward to link it directly to the online version of the coding system. This would also serve to encourage contributions from the wider scholarly community to extend the resource, as this would be of considerable value for future studies of work and labor in the United States during the 19th century.

## Acknowledgments

William G. Thomas, Department of History and the students, staff, faculty, and directors at the Center for Digital Research in the Humanities, all at the University of Nebraska-Lincoln originally provided the raw digitized data file for the Baltimore and Ohio RR 1857 payroll, from which a large number of railroad job titles were drawn for inclusion in the new coding system. Jennifer Mahalidge and Tiffany Rogers coded the mine payroll records from which a number of mining titles were derived.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.



## Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: The support of the ESRC from Grant RES-000-22-2420 "Migration, Economic Opportunity and the Railroads: Movement of Heavy Industrial Workers in the North-East USA 1850-1900" is gratefully acknowledged. The mine payrolls project, from which a number of occupational titles used in the present study were derived, was funded by the Pennsylvania Historical and Museum Commission under the direction of Mary Ann Moran-Savakinas and Ella Rayburn of the Lackawanna Historical Society, Scranton, Pennsylvania.

## References

- Baltimore and Ohio Railroad Company. (1842). *List of persons with their pay in the service of the Baltimore and Ohio Rail Road Company April 1, 1842*. Baltimore, MD: Author.
- Baltimore and Ohio Railroad Company. (1852). *List of officers and employees in the service of the Baltimore and Ohio Rail Road Company with their salaries, duties &c. September 1852*. Baltimore, MD: Author.
- Baltimore and Ohio Railroad Company. (1858). *List of officers and employees of the Baltimore and Ohio Rail Road with the amount of their pay for the month of November 1857*. Baltimore, MD: Author.
- Carter, S. B., & Sutch, R. (1996). Fixing the facts: Editing of the U.S. census of occupations with implications for long-term labor-force trends and the sociology of official statistics. *Historical Methods*, 29, 5-24.
- Chandler, A. D. (1965). *The railroads: The nation's first big business; sources and readings*. New York, NY: Harcourt, Brace & World.
- Commissioner of Labor. (1890). *Fifth annual report of the commissioner of labor, 1889: Railroad labor*. Washington, DC: Government Printing Office.
- Conk, M. A. (1978). Occupational classification in the United States census: 1870-1940. *Journal of Interdisciplinary History*, 9, 111-130.
- Diamond Payroll. (1880). *Diamond mine payroll sheets for July, 1880* (Delaware, Lackawanna & Western Railroad Coal Department Payroll Records). Scranton, PA: Lackawanna Historical Society.
- DiCiccio, C. (1996). *Coal and coke in Pennsylvania*. Harrisburg: Pennsylvania Historical and Museum Commission.
- Edwards, A. M. (1933). A socio-economic grouping of the gainful workers of the United States. *Journal of the American Statistical Association*, 28, 377-387.
- Edwards, A. M. (1943). *Comparative occupation statistics for the United States, 1870 to 1940*. Washington, DC: Government Printing Office.
- Ferrie, J. P. (1995). Up and out or down and out? Immigrant mobility in the antebellum United States. *Journal of Interdisciplinary History*, 26, 33-55.
- Ferrie, J. P. (1999). "Yankees now": *European immigrants in the antebellum U.S., 1840-60*. New York, NY: Oxford University Press.
- Ferrie, J. P. (2005). History lessons: The end of American Exceptionalism? Mobility in the United States since 1850. *Journal of Economic Perspectives*, 19, 199-215.
- Goldstein, S. (1954). City directories as sources of migration data. *American Journal of Sociology*, 60, 169-176.
- Healey, R. G. (2010). *A new occupational coding system for 19th century heavy industrial workers* (Version 1: Iron and Steel, Coal Mining and Railroads). Retrieved from [www.nehgis.org](http://www.nehgis.org) (electronic database accessible from contents page via title page link).
- Healey, R. G. (2011). A full-scale implementation of the NAPP 1880 U.S. Census data set using dimensional modeling and data-warehousing technology. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 44, 95-105.
- Healey, R. G. (2013, June 1). *Miners and mining in Scranton during the Civil War: Preliminary findings from the DL&W Coal Dept. Payrolls Project*. Paper delivered at the Anthracite Heritage Conference, Anthracite Heritage Museum, Scranton, PA.
- Healey, R. G., Thomas, W. G., & Lahman, K. (2013). Railroads and regional labor markets in the mid-nineteenth-century United States: A case study of the Baltimore and Ohio Railroad. *Journal of Historical Geography*, 41, 13-32.
- Herschberg, T. (1976). Occupational classification. *Historical Methods Newsletter*, 9, 59-98.
- Hirschman, C., & Mogford, E. (2009). Immigration and the American industrial revolution from 1880 to 1920. *Social Science Research*, 38, 897-920.
- History of Work Information System. (2013). *History of work information system*. Retrieved from <http://historyofwork.iisg.nl/index.php>
- Hudson Coal Company. (1932). *The story of anthracite*. New York, NY: Author.
- Hunt, W. C. (1909). The Federal census of occupations. *Publications of the American Statistical Association*, 11, 467-485.
- Inspectors of Mines. (1880). *Reports of the inspectors of mines of the anthracite coal regions of Pennsylvania for the year 1880*. Harrisburg, PA: Lane S. Hart, State Printer.
- Knowles, A. K. (2013). *Mastering iron: The struggle to modernize an American industry, 1800-1868*. Chicago, IL: The University of Chicago Press.
- Meyer, P. B. (Ed.). (2004). *The weeks report database* (4th ed.). Retrieved from <http://econterms.net/weeksreport/weeksdoc.htm>
- Miller, A. R., Trieman, D. J., Cain, P. S., & Roos, P. A. (1980). *Work, jobs and occupations: A critical review of the Dictionary of Occupational Titles* (Final Report to the U.S. Dept. of Labor from the Committee on Occupational Classification and Analysis). Washington, DC: National Academy Press.
- Minnesota Population Center. (2008). *North Atlantic Population Project: Complete Count Microdata. Version 2.0* [Machine-readable database]. Minneapolis: Minnesota Population Center. Available from <http://www.nappdata.org>
- Morris, R. J. (1990). Occupational coding: Principles and examples. *Historical Social Research*, 15(1), 3-29.
- NAPP Code List. (2013). *Codes and frequencies*. Retrieved from [https://www.nappdata.org/napp-action/variables/US80A416/#codes\\_section](https://www.nappdata.org/napp-action/variables/US80A416/#codes_section)
- Palmer, G. L. (1939). The convertibility list of occupations and the problems of developing it. *Journal of the American Statistical Association*, 34, 693-708.
- Pennsylvania Railroad. (1851). *Fourth annual report of the directors of the Pennsylvania Railroad Co. to the stockholders, December 31, 1850*. Philadelphia, PA: Crissy & Markley.
- Roberts, E., Woollard, M., Ronnander, C., Dillon, L., & Thorvaldsen, G. (2003). Occupational classification in the North Atlantic Population Project. *Historical Methods*, 36, 89-96.

- Ronnander, C. (1999). The classification of work: Applying 1950 census occupation and industry codes to 1920 responses. *Historical Methods*, 32, 151-155.
- Ruggles, S., Alexander, J., Trent, Genadek, K., Goeken, R., Schroeder, M. B., & Sobek, M. (2010). *Integrated Public Use Microdata Series: Version 5.0* [Machine-readable database]. Minneapolis: University of Minnesota.
- Sarkar, S. (2009, November 12). *Spatial movement of immigrant workers: A comparative analysis between North Atlantic countries*. Paper delivered by E. Roberts at the SSHA Conference, Long Beach, CA.
- Secretary of Internal Affairs. (1877). *Annual report of the secretary of internal affairs of the Commonwealth of Pennsylvania, Part III, Industrial Statistics 1876-7*. Harrisburg, PA: Lane S. Hart, State Printer.
- Secretary of Internal Affairs. (1881). *Annual report of the secretary of internal affairs of the Commonwealth of Pennsylvania, Part III, Industrial Statistics 1879-80*. Harrisburg, PA: Lane S. Hart, State Printer.
- Sobek, M., & Dillon, L. (1995). Interpreting work: Classifying occupations in the public use microdata samples. *Historical Methods*, 28, 70-73.
- Temin, P. (1964). *Iron and steel in nineteenth-century America: An economic inquiry*. Cambridge, MA: MIT Press.
- Thomas, B. (1973). *Migration and economic growth*. Cambridge, UK: Cambridge University Press.
- U.S. Bureau of the Census. (1950). *Alphabetical index of occupations and industries: 1950*. Washington DC: Government Printing Office.
- U.S. Census Office. (1870). *Instructions to enumerators*. Retrieved from <https://usa.ipums.org/usa/voliii/inst1870.shtml>
- U.S. Census Office. (1880). *Instructions to enumerators*. Retrieved from [https://www.nappdata.org/napp/resources/enum\\_materials\\_pdf/enum\\_instruct\\_us1880b.pdf](https://www.nappdata.org/napp/resources/enum_materials_pdf/enum_instruct_us1880b.pdf)
- U.S. Census Office. (1883a). *Vol. 1, Statistics of the population of the United States at the tenth census (June 1, 1880)*. Washington, DC: Government Printing Office.
- U.S. Census Office. (1883b). *Vol. 4, Report on the agencies of transportation in the United States including the statistics of railroads, steam navigation, canals, telegraphs, and telephones*. Washington, DC: Government Printing Office.
- U.S. Census Office. (1890). *1890 census: Instructions to enumerators*. Retrieved from <https://usa.ipums.org/usa/voliii/inst1890.shtml>
- U.S. Census Office. (1895). *Vol. XIV, Report on transportation business in the United States at the Eleventh Census: 1890, Part I, transportation by land*. Washington, DC: Government Printing Office.
- U.S. Census Office. (1900). *Instructions to enumerators*. Retrieved from [https://www.nappdata.org/napp/resources/enum\\_materials\\_pdf/enum\\_instruct\\_us1900a.pdf](https://www.nappdata.org/napp/resources/enum_materials_pdf/enum_instruct_us1900a.pdf)
- U.S. Securities and Exchange Commission. (2011). Division of Corporation Finance: Standard Industrial Classification (SIC) code list. Retrieved from <http://www.sec.gov/info/edgar/sic-codes.htm>
- Vance, J. E., Jr. (1995). *The North American railroad: Its origin, evolution and geography*. Baltimore, MD: Johns Hopkins University Press.
- Van Leeuwen, M. H. D., Maas, I., & Miles, A. (2004). Creating a historical international standard classification of occupations: An exercise in multinational interdisciplinary cooperation. *Historical Methods*, 37, 186-197.
- Weeks, J. D. (Ed.). (1884). *Report on the statistics of wages in manufacturing industries, 1880 Census Vol. XX*. Washington, DC: Government Printing Office.
- Winkler, W. E. (2006). *Overview of record linkage and current research directions* (U.S. Census Bureau Research Report Series, Statistics No. 2006-2). Washington, DC: Bureau of the Census.
- Woollard, M. (1999). *The classification of occupations in the 1881 census of England and Wales* (Historical Censuses and Social Surveys Research Group, Occasional Paper No. 1). Colchester, UK: University of Essex.
- Wrigley, E. A. (n.d.). *The PST system*. Retrieved from <http://www.hpss.geog.cam.ac.uk/research/projects/occupations/categorisation/pst.pdf>

## Author Biography

**Richard Healey** is professor of Geography at the University of Portsmouth in England, having previously taught at the University of Edinburgh. His substantive research interests are focused on the 19th century industrial development of the North-East United States and his technical research interests are in data warehousing, Big Data and database methods in GIS.