The Family History Data Sources Access System (FDAS)

By Richard G. Healey, Ph.D.

Introduction and Background to the Development of the Project

The Family History Data Sources Access System (FDAS for short) is a new web resource in its first stages of development. It is a pleasure to have the opportunity to introduce it to a wider audience. FDAS traces its earliest origins as far back as May 1978 when the present writer first knocked on the door of the Lackawanna Historical Society (LHS) as a fresh-faced postgraduate student working on the development of the 19th Century anthracite coal industry.

Despite the intervening distance, both in space and time, that first contact has grown into a series of long-standing research collaborations with organisations who are now partners in the current project. These organizations include LHS, the Anthracite Heritage Museum, the Steamtown National Historic Site and the School of the Environment, Geography and Geosciences, University of Portsmouth, UK.

The earliest of these collaborations dates back to 2005 when I was a Scholar-in-Residence under the Pennsylvania Historical and Museum Commission (PHMC) scheme, based at the Anthracite Heritage Museum. Further past projects include work on the records of the Delaware, Lackawanna and Western Railroad (DL&WRR) Coal Department, also funded by PHMC, in collaboration with LHS; a project on the migration of heavy industrial workers in the United States during the 19th Century, funded by the UK Economic and Social Research Council; and participation in the 'Digging into Data' Challenge, co-funded by the National Endowment for the Humanities in the United States and the Joint Information Systems Committee in the United Kingdom. This latter project was undertaken in conjunction with the University of Nebraska, Lincoln.

One of the key findings to emerge from the more recent group of projects was that it is not possible to identify accurately the industrial sector to which many skilled workers belonged, solely by reference to the manuscript population censuses in the period 1850-1900. The reason for this is quite straightforward, but its ramifications are far-reaching.

In short, a large majority of what might be termed 'generic tradesmen' self-report as, for example, machinists, carpenters or blacksmiths, but they do not specify their place of work in the census schedules. Hence we do not know whether they are working for a railroad (e.g. in a depot or at the repair shops), in the mines, at a local ironworks, or in some other kind of metal fabricating establishment.

Since this problem affects hundreds of thousands of workers by 1880, any attempt to analyse the *industrial* structure of cities or regions based on individual *occupations* declared in the population census, will be substantially inaccurate, and may lead to serious under-estimation of the economic importance of certain industrial sectors in local and regional economies.

This problem is especially acute in the railroad sector, both because it employed large numbers of men (about 418,000 in 1880 and about 750,000 in 1890) and because a substantial proportion of these workers were generic tradesmen of the kind described above (We know the industry totals in 1880 and 1890 because of special transportation censuses of railroad companies taken in those years. These censuses were conducted independently of the population census). While we may know the total number of railroad men employed across the nation in some census years, we cannot identify a large proportion of them individually from the census alone, so we cannot reliably track their migration movements from place to place or analyse their demographic characteristics in specific locations.

To begin addressing the shortcomings of the 19th Century population census for the study of industrial structure and worker migration within the US, the long-term 'Million Railroaders Project' was established in 2014. This is the most recent research collaboration with Scranton-based heritage organisations, though its eventual geographical coverage is expected to be much wider than northeast Pennsylvania.

Its primary research aims are first, to locate and utilise non-census data sources on industrial employment, which are sufficiently detailed to identify the company affiliations of workers, such that individuals can be correctly classified by industrial sector. Second, wherever possible, using available printed works or online public domain sources (but not online subscription services), to link the resulting employment records to individual census records. Third, to analyse the resulting sets of data records, linked between different data sources and to the census, to begin the complex process of tracing the movements of workers, not only between industrial sectors, but also between locations over time.

The primary research focus of the 'Million Railroaders Project' drew its inspiration from historical geography, and economic and labour history, rather than genealogy. However, as the project concept began to take shape, it quickly became clear that many of the resulting data resources needed to answer broad research questions would have potential additional value, at the detailed level, in relation to a wider

public interest in family history, although it would take both time and considerable effort for this potential to be realised.

Several years later, and after significant data collection and processing had taken place on the Railroaders project (which had also started to widen its scope to include both mine and iron/steel workers), the time arrived to begin harnessing the previously identified potential for family history research. So the lengthy process of developing an online system, named FDAS (at least for the present!) commenced, to provide public access to selected subsets of the available Railroaders project data.

Overview of Key Distinguishing Characteristics of FDAS

Based on the foregoing explanation, the first key difference between FDAS and subscription-based family history sites is that the project is rooted in academic research, not commercial considerations, and has no connection with any other online service providers. Second, since the origins of the work were not in the field of genealogy, as noted above, the project is best viewed as a complement to, rather than a replacement for currently available family history resources on the web. Third, it is anticipated that the fully-fledged version of the system will contain a variety of data not available elsewhere, as well as other data sources, which will be better known to family history researchers, but which may be presented in different and sometimes more flexible forms than the versions accessible from other web sites. Fourth, the original research focus on heavy industrial workers necessarily means that the data to be made available will be more *selective* than those found on general purpose genealogical sites, hence the earlier point about it being complementary to the latter.

The other side of the coin, however, in this regard, is that individual data records may be rather more informative than those found on other sites for reasons that will be examined further below. Finally, since the data resources are managed by the project, rather than a third-party commercial provider, it should be possible, over time, to begin developing what might be called 'value-added' educational and public information resources derived from the 'raw' data in response to requests from groups of users.

Planned Digital Data Resources

The first group of data resources comprise those that will be familiar to all family historians. They include:

- City Directories
- Manuscript Population Census Records

City directories are of particular value in the present context, as they often contain more information about individuals than census entries and sometimes very specific job titles, such as 'Blacksmith in Car Shop' followed by a specific railroad name. Instances such as this are extremely useful, since they remove any doubt about the industrial sector and workplace location, while also confirming the more generic occupation title. However, the level of detail in individual directory entries can vary markedly both within and between directories published in different years for the

same town/city, never mind between directories for different locations. In general, directories published in the late 1850s and 1860s have more details on individuals than later directories, although there are important exceptions to this rule.

The main shortcoming of city directories is the obvious lack of any useful data on family status and family members, since usually only 'heads of household' and unmarried offspring of working age tend to be reported. The less obvious absence of data on age and birthplace is problematic because these characteristics are very helpful for matching individuals to census entries.

In comparison, it is well-known that the manuscript population census schedules contain a wealth of demographic and economic detail about individuals and their families, which increases in volume during the 19th Century, so this does not need to be described at length here. They also have a claim to comprehensiveness that other sources lack, although the extent of under-reporting in rapidly developing industrial regions prior to 1880 may well have been under-estimated, in the present writer's view.

Since these census schedules are readily available elsewhere on a subscription basis, this project will not attempt to duplicate the enormous investments already made to make comprehensive census resources available to the public. That said, the matching process between census and other records, outlined earlier, will require a growing body of selected records, where matches have been found, to be included in the present system. The schedules used are, and will be, those provided either by scans of census microfilms available on the public domain Internet Archive, or from original reel copies purchased directly from the National Archives and Records Administration.

Examples of industrial sectors where city directories may prove helpful, in the absence of other types of industrial records (see below), include the iron and steel industry and the oil industry. Contemporary photographs from the author's collection convey a sense of the types of workplace involved:
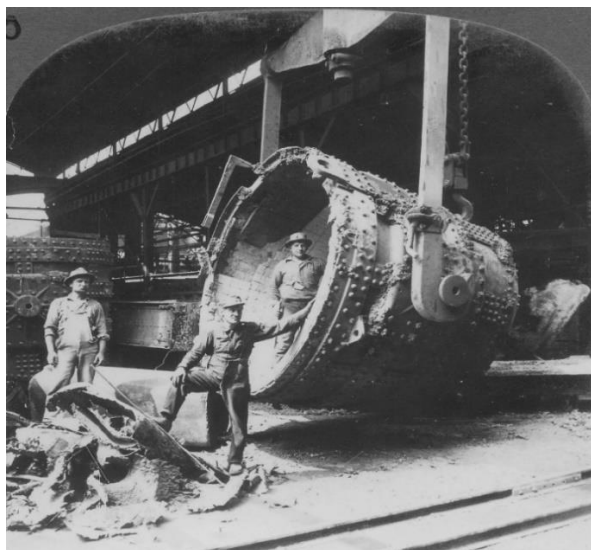
Cleaning and Repairing a Ladle for Molten Steel – Pittsburgh:
Keystone View Company photograph

Troutman Well – Pennsylvania Oil Regions: Frank Robbins, Photographer

The second group of resources comprises more specialised datasets that will still be familiar to some but not all family historians. They comprise:

- Civil War Draft Registration Records
- Mine Accident Records

The Civil War Draft Registration Records are a very large collection of manuscript volumes listing details of men between the ages of 20 and 45 eligible for military service during the Civil War. The overall organisation of subsets of volumes is by congressional district and within those subsets by different categories of individuals.

Within each individual volume, entries are only arranged by the first letter of each person's name, so to find all the individuals in a specific local area such as a township, would require an inordinately time consuming search through tens or even hundreds of thousands of entries spread across hundreds of manuscript pages in multiple volumes.

Needless to say, providing access to the totality of these records is far beyond the resources available to the current project. Instead, convenient search access to data for a small number of Pennsylvania and Ohio counties is planned in areas where the system will have good coverage using other records, but at different dates. The advantage of the Civil War Draft Registration Records is that, like the census, they contain age, occupation, marital status and birthplace, which facilitates possible matching to the census, although no data on families are provided.

The Civil War Draft Registration Record data are made available with the specific permission of the National Archives and Records Administration, Washington, DC. Mine accident records for the Anthracite Coal Regions are found in the Anthracite Mine Inspectors' Reports published after 1870. These records are quite well-known to family historians in northeast Pennsylvania, but less well–known elsewhere. Similar records are available for the bituminous coal regions of Pennsylvania and Ohio. All the records contain details of the accident itself, the name of the employee affected, the mine involved and usually whether the accident was fatal. However, in the 1870s the majority of records lack data on the age of the employee, which makes matching to other records more difficult. This situation improves in the 1880s.

Initially, the system will provide access to selected accident data originally made accessible in a stand-alone fashion by the present writer some years ago (see www.nehgis.org for details). This is simply for convenience when working with multiple sources, but additional data not previously available will be added.
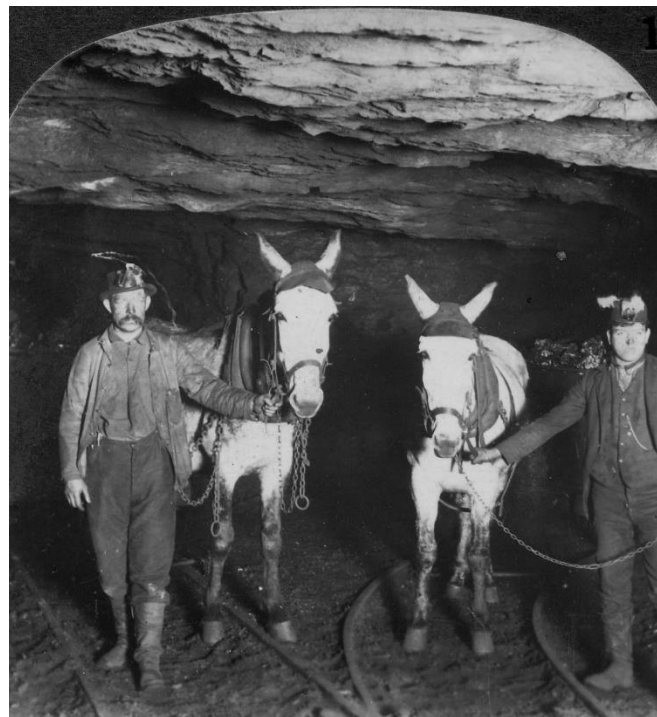
Traditional image of miners and mules underground:
Keystone View Company photograph

The third group of resources have not previously been made available on the web. They comprise:

- A name index to the Civil War Era Payrolls of the DL&WRR Coal Department
- Railroad Accident Records

The payroll name index covers the period 1860-1864 and indicates both the date of first employment of a specific individual in the payrolls and the mine where that employment commenced. Individuals already at work at the start of the period covered are simply given a 'start date' of January 1860.

As students of railroad history will be aware, data on railroad accidents can be found scattered across an immensely wide range of 19th Century publications, including newspapers, company annual reports and state railroad commission reports, often spanning extended periods of years. As with mine accident data, the level of detail also varies considerably both within and between reports, and often, though not always, later reports from the 1870s and 1880s can be more informative than earlier ones. As usual, the approach of this project is to identify and process some of the better quality records, first beginning with a focus on upper New York State and spreading outwards to adjacent states.

The final group of planned resources are not yet implemented in FDAS, although some limited digitisation has already taken place and a range of source materials have already been identified for future digitisation. They currently comprise:

- Railroad Report Employee Lists
- Railroad Payroll and Non-Payroll Employee Lists

Employee lists in railroad reports are arguably even more scattered than railroad accident records as they are not a systematic and standardised feature of annual reports to stockholders. Nevertheless, in a number of cases, railroads either listed key employees annually for a series of years, or, as in the case of the Baltimore and Ohio Railroad, published for a few selected years, lengthy lists of employees.

Although such records have little or no information that is immediately useful from a demographic perspective, they have the obvious and key advantage of unambiguously identifying individuals actually employed in particular occupations by a specific railroad in a specific year. In certain cases, such as with depot agents, the location of that employment is also given. In combination with other sources, records of this kind may enable workers only listed as machinists, in a city directory for example, to be more accurately classified as also railroad employees.

To some researchers, manuscript railroad payroll records are an elusive 'holy grail', given the extraordinary paucity of such records in comparison to the very large size of the railroad sector by the latter part of the 19th Century. However, occasionally, other related, but non-payroll sources, such as railroad disciplinary records, strike records, and employee lists (or black-lists!) in company correspondence, may provide additional limited insights into workers at specific dates and locations. Over the course of forty years of research into 19th Century US industrial development, the present writer has accumulated some interesting material of this kind, and it is the intention to make it available progressively through the FDAS system.
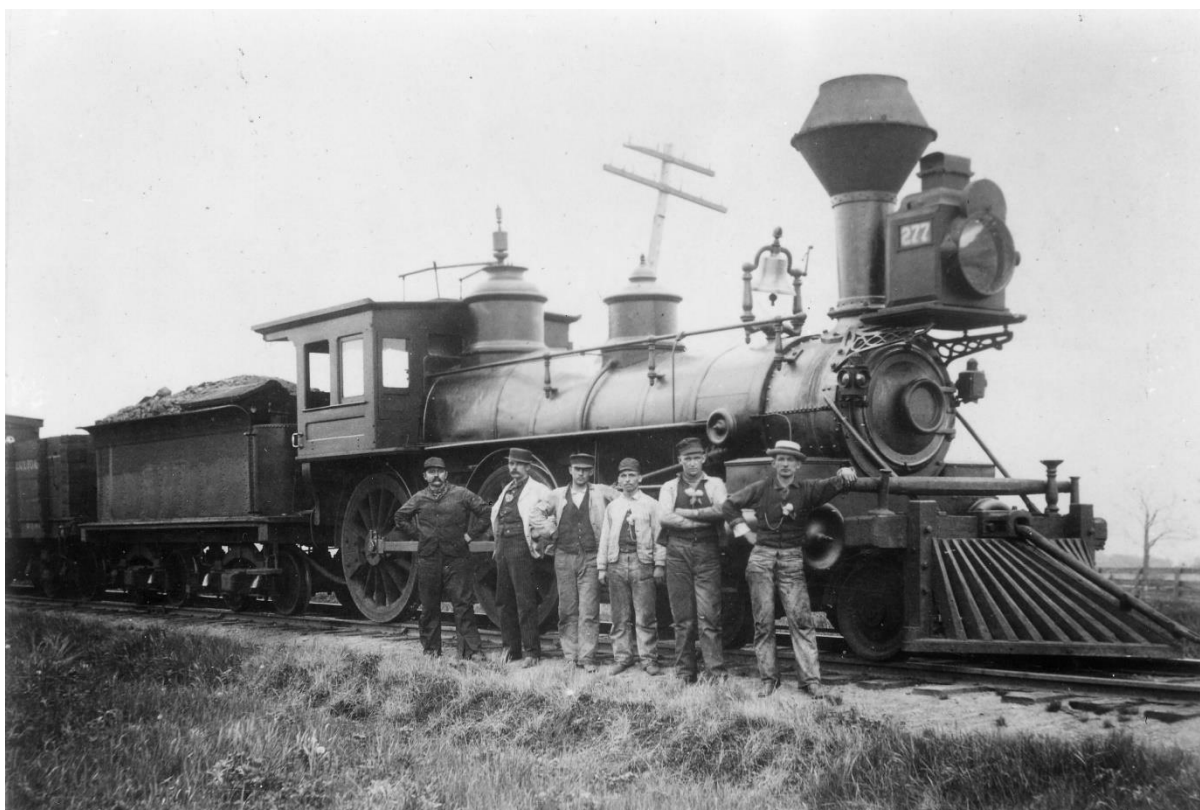
The above image is a relatively rare early example of a train crew photograph where a number of the employees are identified. This photograph was taken on the New York Central and Hudson River Railroad circa 1878 near Schenectady – contextual information that increases the likelihood of finding further references to the individuals in question in city directory or railroad accident records. From left to right, the workers are as follows:

Charles H. Smith: Engineer; Seymour Conway: Conductor; Bert Lytell: Fireman; Unidentified Employee; William Flanagan: Brakeman; Unidentified Employee.

Temporal and Geographical Coverage in FDAS

The main time period for many of the current and planned resources in FDAS is between 1850 and 1880 although some datasets will extend beyond this, both earlier and later. Data from before the start of the 'railway age', around 1830 or much after 1900, will not be a priority for the foreseeable future. In contrast, a particular early focus will be on the Civil War era, taken here to be from about 1855-1870.

In geographical terms, it is very appropriate for the first public announcement about the system to appear in *History Bytes*, as the Anthracite Coal Regions can be regarded as the initial core area, since the first group of partner organisations are based in Scranton. Over time, the geographical coverage will expand approximately along the lines of the anthracite railroads and canals – south towards Baltimore and Philadelphia, east via the Lehigh Valley in the direction of the New Jersey manufacturing centres and New York City, north towards upper New York State and westwards along the Pennsylvania Railroad across Pennsylvania and eventually into eastern Ohio.

Thus, taking the example of city directories, the initial public version of the system will include Scranton directories 1859-1875 and early Wilkes-Barre and Hazleton directories. Subsequently, early directories for Allentown, Easton, Bethlehem, Pottsville, Reading and Williamsport will be added and the boundary of geographical coverage will continue to expand thereafter.

Avondale Mine Disaster 150th Anniversary Commemoration

In recognition of the Avondale Mine Disaster 150th Anniversary Commemoration that was held in September 2019, a special category will be added to the Mine Accidents section of FDAS to record the names of the men and boys who perished in the disaster. While lists of the names are already accessible elsewhere on the Web, making them directly available within the system will make it easier for researchers to combine the names with other related searches on different sources, if desired.

Further to this, though it is a longer-term undertaking, priority will be given to the inclusion of Civil War Draft Registration records for townships in the vicinity of Avondale to facilitate a greater appreciation of the employment context of the area in the years preceding the disaster.

System Implementation

The system is hosted by the University of Portsmouth in England on a powerful server with enterprise class cyber-security measures in place. It currently uses the world-leading ORACLE™ database software to provide a web-connected database service. It is intended to have high availability.

Unlike many subscription services and online genealogical sites, the system does not aim to serve up images of original documents based on surname indexes, built either manually or using OCR software (more or less effectively). Rather, all the entries selected from a given source are digitised, double-checked for accuracy and coded with standardised occupational codes and, where possible, standardised, corrected work location names (e.g. mine names), to facilitate reliable searching on employment as well as demographic characteristics.

Users should therefore have an expectation that *at least* 99% of entries will accurately reflect the content of the original and in some cases the records will contain enhanced information compared to the original. For example, while the name of a DL&WRR Coal Department mine may be misspelled in the original, the correct names are known to the system and will be reflected in the output provided to the user. This approach has been adopted because a very high degree of accuracy is required for database records that are also to be used for research analyses, and the benefit of this can then be passed on directly to FDAS users.

It should be emphasized that a selective approach to data collection means that, wherever possible, all workers who are definitely or potentially (the unspecified generic tradesmen) employed in the heavy industrial sectors of interest will be included, but others working in retail trades, for example, will not. Hence on average, one quarter to one third of all the entries in a given directory may be included, only very rarely will the fraction drop below one-tenth or exceed one-half.

Another beneficial side effect of not providing images is that searches are extremely fast and can easily allow for more complex requirements, such as searching for repeated occurrences of individuals across city directories that were published in different years.

Stages in the FDAS Development Process

There are four main stages in the initial development of FDAS:

- Development of a Version 0 concept system
- Testing of the concept system for operational or user interface problems
- Beta testing of Version 1 of the system and 'bug fixing'
- Launch of Version 1 of the publicly available system

The concept system with a 'bare bones' user interface was developed in the latter part of 2018 and successfully tested at a workshop organised by LHS and kindly hosted at the University of Scranton in September 2018. The enthusiastic participants provided valuable feedback both on system configuration and priorities for further dataset inclusion.

Since then further dataset development has taken place, with more to come, and attention has turned to the establishment of a beta testing programme over the coming months, probably after a major upgrade to core database software has taken place. The original workshop participants will be invited to act as beta testers, but there is scope for a limited number of additional testers to join them.

It would be of particular interest to hear from possible testers who could provide input from the public library and the high school education sectors, but others from different parts of the heritage sector are also most welcome. The beta testing programme is open to individuals or representatives of non-profit organisations.

If you would like to take part, please contact me by email (address at the foot of the article) stating the reasons for your interest, and ways in which you think the system might help your personal family history research or the mission of your organisation. All participants will be asked to complete free registration on the system.

Once the beta testing process has been completed, the feedback will be utilised to finalise the first production version of the system for public release. The date of this release will be announced in a subsequent issue of *History Bytes*.

The present writer is very happy to respond to email questions about the system itself, but, to use the customary caveat on these occasions – is not able to answer specific genealogical/family history questions, which should, as always, be directed to the appropriate libraries or heritage organisations.

Richard G. Healey, Ph.D.

School of the Environment, Geography and Geosciences
University of Portsmouth, United Kingdom
Email: richard.healey@port.ac.uk