Collocation is an example of a digital humanities approach suggestive of links and connections that require further research and analysis. The close reading of texts, as modelled in this first part of this article, is an important step for understanding how physicians, newspapers and, indeed, the public understood this epidemic in its earliest stages. Yet text visualisation tools also have great potential to identify important themes, to suggest connections and to identify possible relationships. Researchers can pursue this research on their own using a combination of full text sources in the Medical Heritage Library, text visualisations in Voyant tools and network analysis in Palladio. For students working on medical research projects, these tools combine relative ease of access and applications with the possibility of increasingly sophisticated analytical strategies that yield new insights. This approach allows scholars as well as students to appreciate the value of a digital humanities approach as well as the importance of close-reading skills to explore more fully the historical significance of an event such as an influenza epidemic.

E. Thomas Ewing
Virginia Tech, Blacksburg, VA, USA

doi:10.1017/mdh.2017.54

New Methods in the History of Medicine: Streamlining Workflows to Enable Big-Data History Projects

This paper presents new methods, workflows and a project management system that we developed to reduce the resources needed for big-data history projects and thus lower the barriers to entry for other scholars. Creating datasets from handwritten documents – essentially constructing a new archive from which to investigate historical questions – shifts the traditional timeline and resource requirements of historical research. This is a double-edged sword. Once the dataset is built, researchers can use it to investigate a wide range of questions. Yet, building a dataset requires a substantial investment of resources (i.e., knowledge, time, labour and money).

We developed these new approaches, out of necessity, for the New Orleans Mortality Project (http://nola.spatialhistory.org), an interdisciplinary historical geographic information systems (HGIS) study of the impact of disease, socio-economics and environment on community and urban development, 1877–1915. First, this paper details the workflows we developed in order to build a 50 000-record mortality database from death certificates, a 40 000-record property value database from tax ledgers and city-wide population datasets from city directories. Second, the paper explains the project management system we created to foster efficient and accurate database creation by undergraduate students. Developing these workflows and project management techniques made the large scope and depth of the project possible. Third, this paper presents the results of this project management approach and discusses the broader implications of these findings. Methodological innovations and lessons from this project can be incorporated into a large variety of other digital history projects.

Like many nineteenth-century administrative records, the state and city death certificates and the Orleans Parish Assessor's records presented two challenges for digitisation: script handwriting and a variety of hands (from different recorders). Advances in optical character recognition (OCR) continue to unlock historical records for further analysis; however, OCR remains severely limited when working with script handwriting. Extensive

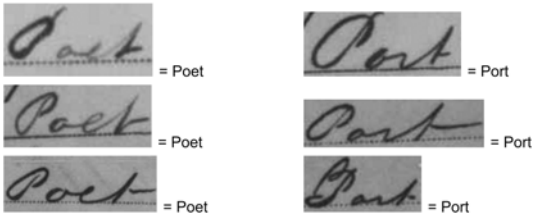

Figure 1: Training document illustrating the difference in handwriting between Poet St. and Port St.

pattern training is often required for accuracy, and trained patterns rarely work for different handwritings. At present, building databases (structured datasets) from these records cannot be automated. Therefore, we hired undergraduate research assistants and developed methods for these students to work with the records accurately and efficiently.

This project required an online (cloud) platform to facilitate collaborative database development, and we chose Google Drive for two reasons. First, Drive has the most seamless online collaboration options in Docs, Sheets and Forms. This meant that the whole project team could work on a single spreadsheet at the same time, even when students were spread across the country. With an internet-connected computer, students had access to all the training documents, the death certificate images and the spreadsheets. Second, Docs and Sheets allow for export to non-proprietary formats, such as plain text (.txt) and comma separated values (.csv). This helps to minimise the risk that the data will end up inaccessible in an obsolete format or defunct service.

Training students to interpret the mortality records and enter the data efficiently was key to building the mortality database. We created tutorial videos for new students (sample training video: https://youtu.be/fKjbkWJGMvE). This reduced the initial training burden, allowing us to quickly hire and train more students, and it freed us to focus on what the students did not understand.

We were able to isolate the records and handwriting the students could not read by creating an 'Attention Needed' field. Instead of spending lots of time trying to interpret each difficult record, students could simply flag problematic records. This method let students focus on the records they could interpret, and it allowed us to quickly resolve many of the issues marked by students since we had more local knowledge of New Orleans and a familiarity with historical diseases. Furthermore, we shared this knowledge with students by creating training guidelines based on the recurring issues marked in the 'Attention Needed' field. We assembled guides that showed multiple screenshots of commonly confused words and numbers so students could start to learn the subtle differences in the handwriting, such as the way a recorder wrote 'Port Street' versus 'Poet Street' (see Figure 1).

Another important contribution is our method of data validation. We needed to validate and standardise the student entries in important fields such as Nativity, Street Name, Cause of Death and Occupation, so we created controlled vocabulary lists. These lists were stored in a separate spreadsheet and linked to the corresponding mortality spreadsheet fields through the *Data Validation* function in Google Sheets. This function added a dropdown menu to the spreadsheet fields with a list of possible entries that were filtered after each keystroke by the student. In training, we instructed students that if they were not familiar with the cause of death or could not interpret the entire word on a death certificate, to type the first letters they could read. For example, a student not familiar with Phthisis

https://doi.org/10.1017/mdh.2017.54 Published online by Cambridge University Press

Pulmonalis (i.e., tuberculosis) only needed to type 'PH' into the cause of death field, and Phthisis Pulmonalis appeared as one of the choices in the dropdown list.

We populated these lists with data from three sources. First, we entered modern and historical street names. Modern street names were from the US Census TIGER files, and the historical names were from a 1912 street guide and a 1938 report published by the New Orleans City Hall Archives. Second, we entered the causes of death listed in the local Charity Hospital annual reports ranging from 1880 to 1915. Third, we added the entries from 2500 records (5% of total) we entered in the testing phase.

To further optimise this system, we created a live vocabulary list that pulled the data entered by students, in real time, from all the mortality spreadsheets using the *ImportRange* function. The new spreadsheet sorted the entries based on frequency counts from the mortality spreadsheets rather than alphabetically. We linked the new spreadsheet to the original vocabulary lists and built filters that limited the data pulled into the original to only approved entries, which we regularly updated. This allowed us to review the new entries and prevent incorrect entries from populating the controlled vocabulary lists. The frequency count ranked the entries in the dropdown list so that, building from the previous example, after typing 'PH' in the cause of death field, Phthisis Pulmonalis appeared first in the list because it was the most common cause of death in the records, and uncommon causes of death, like Phosphorus Poisoning (an alphabetical antecedent), appeared towards the bottom of the list.

Creating and populating the controlled vocabulary lists took about thirty hours of work upfront (not including the 2500-record testing phase), but they ultimately saved much more time than they cost. The lists standardised the field entries, and also helped to familiarise students with the common streets, place names and causes of death listed in the records. Moreover, these lists saved time by reducing the number of manual keystrokes needed for each entry and also significantly reduced errors; we tracked both improvements in the project management system.

We developed a project management system to monitor real-time efficiency, project budget, completion status and accuracy. Students submitted a Google Form timesheet for each work session, which included their start and end times, the start and end rows in the spreadsheet, the number of Attention Needed records, and any notes or issues. The form responses populated a spreadsheet, which we set up to calculate the number of records completed, the records per hour for each work session and the student pay (not connected with performance). We sampled the accuracy of student work sessions and logged the error statistics next to each work session in the spreadsheet.

We used the *ImportRange* and the *CountBlank* functions to generate status information for each mortality and tax spreadsheet. This calculated, in real time, the number of incomplete records. This number was divided by the total number of records to provide statistics on the completed percentage of the database. Since the handwriting in different years affected the efficiency, we created another field that filtered the student efficiency data by year. From this we calculated an HoursRemaining field, which predicted the number of hours remaining for each year and for the complete mortality database. The ability to predict, down to the hour, the time to completion proved very beneficial for project planning, managing the number of employees, scheduling trips to the archive to scan more records and maintaining project momentum in general.

We also calculated descriptive statistics for the work done by each student. Students had access to their own performance data and statistics, and we emphasised that the efficiency rates and accuracy checks were not a grade, but instead a tool to identify issues and provide

https://doi.org/10.1017/mdh.2017.54 Published online by Cambridge University Press

more specific training. The efficiency and accuracy statistics helped us identify and remedy bottlenecks in the workflows. Finally, we charted the total records completed each week, the total money spent and predicted the total records that would be finished the week when the grant money ran out.

As a result of these workflows and project management system, we found large improvements in student efficiency and accuracy. The resulting database includes 50 000 mortality records for every fifth year between 1880 and 1915. The tax database has property values for about 5000 blocks for each of the eight sample years (40 000 block values in total). After building the mortality and property tax databases, we built historical address locators to link each record to the residential address in the HGIS.[1] We reviewed 20% of the total records entered and the accuracy of these records was above 95%. These new methods resulted in an approximately 50% reduction in cost and completion time. Students averaged 40–60 records per hour without these enhanced workflows and 80–120 records per hour after implementation. Average cost per mortality record (15 fields) was $0.1278 compared to commercial data-entry rates of $0.1825 per record.

This paper has detailed the system developed for creating and documenting the analog-to-digital data creation process in the New Orleans Mortality Project. Historians can easily apply parts of our workflow, such as using terminology libraries for data validation and creating issue-specific training documents based on flagged records, to streamline construction of datasets. Moreover, the project management system we developed can be used in an even wider range of research projects to manage students and budgets, identify bottlenecks in workflows and for planning purposes. These techniques substantially reduced the resource requirements to create structured datasets from handwritten documents. In turn, the detail and extent of our dataset allows us to answer historical questions in new and exciting ways.

Acknowledgements

The New Orleans Mortality Project has received support from the Rice Humanities Research Center, the Rice History Department, the Economic History Association and the Federal Work-Study Program.

S. Wright Kennedy, Jessica C. Kuzmin and Benjamin Jones
Rice University, Houston, USA

[1] For examples of historical address locators, see Peter Tuckel, Sharon Sassler, Richard Maisel and Andrew Leykam, 'The Diffusion of the Influenza Pandemic of 1918 in Hartford, Connecticut', *Social Science History* 30, 2 (2006), 167–96; Don Lafreniere and Jason Gilliland, ' "All the World's a Stage": A GIS Framework for Recreating Personal Time-Space from Qualitative and Quantitative Sources', *Transactions in GIS*, 1 July 2014.

https://doi.org/10.1017/mdh.2017.54 Published online by Cambridge University Press