# Bodies in Transit Project Progress Reports

1.

My group's project with the New York Municipal Library is centered around their Bodies in Transit records, which detail the passage of corpses through NYC ports from 1859-1894. Because these records were kept by the city's Department of Health–ostensibly in order to track disease–the information listed includes the date of passage, name, age, nativity, place of death, cause of death, date of death, and place of interment of the body, as well as the name of the person given charge of the body. There are also columns for residence of the deceased, occupation of the deceased, and miscellaneous remarks, although these are filled in less diligently. The records are listed chronologically and alphabetically within each time period.

Given this information, we have visualized a few ways in which to represent the data for our final project. A recurring idea, and perhaps the most sensical, is to map each person's passage on a map, our first plot point marking where they were born; then where they lived, if disclosed; where they died; where their body was interred. Professor Wolf commented that it would be interesting to see each set of data converge in New York; interestingly, some people's details elude New York City completely, but a majority of records are tied in at least one way to the Northeastern United States. We will have to tinker with what we want to demonstrate and what will prove the most interesting points. For example, if there is a way to actually track disease geographically, that would be relevant to the Department of Health's initial interests, but may not be the most effective use of all our data.

Our main challenge at this point is the sheer amount of data we have, and how comparatively little time we have to dedicate to this project. Even though we have only scanned the first of three ledgers, we have several hundred hours of work on our hands. We are scanning PDFs from our microfilm reels and transcribing said PDFs into a spreadsheet concurrently, but even calling in Professor Wolf as an auxiliary would leave us far short of completing even one ledger by December. We are currently trying to decide what selection of records we should work on, bearing in mind that a) we want to have as representative an amount of data as possible, across all letters and b) we're wary of including only some of each letter, lest the Municipal Library find their information scrambled if they choose to work on this further.

For now, we will keep scanning and transcribing our records. One of my fellow group members will elaborate and update on these matters, as well as transcribing issues, in next week's post.

- Rubi Mora

2.

Our primary subject of conversation this week was (and continues to be) that of data sampling, which Rubi introduced in last week's summary: in a collection that consists of three microfilm reels of ledger pages, how might a project team decide what data to foreground given a limited timeline? On a general level, we chose to constrict ourselves to the ledger's first volume, spanning the years between 1859 and 1864. The reasoning for this decision is twofold: (1) it enables a relative ease of digitization, enabling both us and the hypothetical teams that follow to proceed in sequential chunks, and (2) it permits us to explore the question of the ledger's purpose and creation.

However, when this question of sampling recurs lower in ledger's structural hierarchy — e.g., within the volume — the solution seems less clear. Digitizing and transcribing ledger pages sequentially would ultimately allow for easy transferability: given a clearly-marked stopping point, a future NYU project team member or Municipal Archives representative might simply continue forward in the reel and transcription spreadsheet. The results of this process, however, might prove to be less interesting in the context of applied methods. Situating the end of the project timeline within the span of the Fall 2014 semester produces a different inclination than considering the project to have an indefinite timeline, potentially augmented by students and archivists.

In the former scenario, more consistent data might be extracted by sampling across letters through a consistent, predetermined period of time. The ledger is organized in such a way that all individuals with last names beginning with a certain letter are grouped together for the entirety of the volume's period, so that we see all of the records of individuals with last names beginning with "A" for all of 1859-1864 before the appearance of an 1869 "B" record. To sample data from the beginning of each letter-section for the first, say, two years of the volume's span, would produce a significantly larger data set concentrated in a time period that would then become the focus of study. Mapping routes of corpse movement would also become less tenuous, given that the data points would coincide temporally and be free of the alphabetical representation bias. Such an approach to transcription*, however, would add considerable difficulty to the project's continuation, given that our team would produce difficult-to-locate (as anything in microfilm is difficult to locate) gaps rather than providing a continuous database to which addition could be seamless.

Although we have had some difficulty reaching our contact at the Municipal Archives, as a team with archival slants, we have decided to proceed sequentially. Although the alternative might produce interesting results for us, it would do a disservice to the archive and to future project teams. Because any visualizations or statistical analyses on our part are necessarily discrete and prototypical, we can choose to treat the alphabetical sequential data as a reasonable representation for the purpose of applying methods and technologies. Ultimately, we must be aware of the limitations that our chosen transcription approach produces for visualization and research. Although we will have data from all years of the ledger, which would have been impossible via alphabetical sampling, exploring our primary interest of cause of death over time suffers from the bias that the method introduces. Certainly, we can extract some sense of common causes of death and situate them within changing historical conceptions and naming conventions of disease, as we intended to do, with the caveat that the data set is always partial. A wider range of temporal data will be available, making possible a cause-of-death timeline or slider demonstrating changes and trends between 1859 and 1864. As a result of our decision, we will also have less data from the beginning year of the ledger's existence, which may or may not influence a discussion of the impetus for its production; much of our research places these Department of Health records within the context of sanitary and public health reform in New York City and elsewhere in the U.S. As we move into a research-driven phase of the project, we will continue to think about the ways in which data sampling methodologies influence both research questions that can be asked and the results that might be produced.

*To answer Nick's question about the headache of transcription, the ability to use a collective, real-time document is helpful in facilitating cross-checking and support. I personally have worked and am working on digitization more heavily because I've had such trouble with transcription; aside from being time-intensive (as is transforming microfilm into PDF), it poses a lot of interpretive problems for me as someone who is horrible at deciphering handwriting. The Google Sheets comment function seems to be serving us well, as any team member can flag a certain cell or section of the document with a question or comment, and Drive will push the comment to the rest of the team via email.

- Grace Afsari-Mamagani

3.

Our group has been working tirelessly to transcribe the “Bodies in Transit” collection for the New York Municipal Archives. As end of the semester looms near, we found that we had not transcribed as much as we naively thought we could. This week we had to consider the type of data we would receive from only transcribing the early pages of the ledgers that were in alphabetical order. We have been in the middle of deciding if we should progress sequentially or sample from the data in order for us to get a more diversified result, we finally concluded sequentially makes the most sense in the long run. Even if we are not the people to finish the data set, it makes it much easier for another group to pick up where we left off. While this may lead us to have less diversified data, it is a smarter choice. We set some deadlines for ourselves in terms of transcribing so that we would have time to clean up the data. While we will not be the last people to touch the data, our work will help to create a standard for the next transcribers so that all the ledgers will be uniform upon completion and become quality data.

The most difficult aspect of having four different people transcribing handwriting into database is the lack of consistency. We were nervous about that since our results frequently showed that some of us simply put “Brooklyn” for a location, while others put “Brooklyn, New York.” Sometimes it may be more obvious to someone that an entry for place of death states “Bridgeport, Conn” meaning Bridgeport, Connecticut due to previous knowledge or the way they interpret the hand writing. We want to make sure the data is a clean as possible, and that is where we will develop our “standard” for transcribing the records for the next group so that it is consistent.

Just this week we met and put the data into Google Open Refine and were pleasantly surprised that our data was relatively clean. In the end this will be one of the things we clean up within the dataset so that we can have more relatable data to spot trends. As a way to combat the struggle that is transcribing we have worked together in a shared document to record common terms that came up that were not self-explanatory. When we work with people’s names, town names, causes of death, and place of burial there as similar items. While it is it easier to transcribe Typhoid Fever because it is something that we are familiar with, it is much more difficult to transcribe a word we have been previously unaccustomed to such as “Marasmus.” Upon research, marasmus is just another term for consumption, which was a blanket term for the body failing due to sickness. It is important and useful to take control of the unfamiliar data so that we record it uniformly to analyze later on.

Our transcriptions will end on December 2nd and our last phase of the project will be visualizing the data. We have worked with our raw data in CartoDB and may go back down that avenue once we clean it up.

- Shannon McDonald

4.

With our transcription completed, the final stage of our project discusses what to do with all the data.

Even though we only transcribed a fraction of the records, we were still faced with numerous possibilities for interpretation and visualization. But before we could move ahead we had to clean our data. As Shannon mentioned earlier, there were numerous inconsistencies in our transcriptions. This was a result of various people transcribing but also inconsistencies in the ledger itself. The most common inconsistencies were in place names and cemetery names. The data cleaning was more complicated than we expected because of the various spellings and transcriptions.

Once the data was scrubbed, we turned our attention to different visualizations that we could employ. There were a few things we had to consider before we moved forwards with our visualizations. With only a small portion of the data transcribed we had to decide what was going to accurately represent our data, examine its significance, and highlight the areas of the record that were most compelling.

Through conversations with the Municipal Archives at the inception of the project, we thought mapping the route of the bodies – origin, New York, destination – would be a great visualization. This was easier said than done. The biggest challenge we faced with mapping was the inconsistencies in the record. Many entries simply didn't record where the body originated from which made creating a three point map with those entries impossible. Despite the inconsistencies in some of the records we were still able to generate two maps that documented the route of the bodies. The first map, used 811 records with noted place of death and internment, showing the paths in and out of New York City. Using 469 records, the second map showed the corpses paths between nativity, place of death, New York City and internment.

We also wanted to highlight other significant areas of the record. We chose places of internment and cause of death, as two other aspects of the record to visualize. As we worked with the data we noticed the popularity of certain cemeteries in and around the New York area. By using two different types of graph visualizations we were able to show the top cemeteries where people were interred. We were also able to show the relationship between place of death and cemeteries.

The final area we wanted to focus on was cause of death. Through our visualizations we highlighted the multiple causes of death over the time periods. We wanted to show the overall causes of death and the percentage of entries in the record. We provided further visualization by displaying the data by year and highlighting the soldiers documented in the record.

When we were discussing different visualization models for our data set, we tried to emphasize the potential for the complete record. We wanted our visualizations to demonstrate what the data could contain for the Municipal Archives and future researchers and what it could contribute to the history of New York City.

- Victoria Harty