Backend Visualizations:
Tools for Helping the Research Process

Ryan Delaney[1]

Most visualizations represent the end point of a specific line of inquiry within a project; a map in a paper, an animation telling a previously-determined story, or an interactive mechanism for browsing a project's findings. Visualizations however, need not be restricted to presenting 'complicated' information to lay-people, but can, in fact, rapidly increase the productivity of any given task, from data entry to helping set the research direction. In this paper, we demonstrate two such utilities that we have developed, along with their methods for visualization, which have helped us in at least a twofold manner; each was developed and used for error-checking, but then helped to steer the course of our further research, once their potentials were unlocked.

While visualizations may typically be thought of, and frequently employed, as end-products to help convey a previously identified story or idea, a visualization may also be utilized to help quickly identify anomalous data to either correct errors or help direct further research. As part of the Terrain of History project investigating the 19th century history of the Brazilian city Rio de Janeiro, we have developed two such quality control programs that have helped us not only to clean our data but also to direct us towards further areas of research. Before delving too deeply into these utilities however, it is necessary to give a brief project overview, in terms of sources and data.

Background

The Terrain of History project is investigating the social and economic history of Rio de Janeiro in the latter half of the 19th century and to that end we have a multitude of datasets, many of which are quite expansive, touching on numerous aspects of life in the city during that time. Among our datasets, and most importantly for the purposes of this paper, there are three key ones: the Decima Urbana (ARCRJ), Almanak Laemmert and Voter Qualification Rolls (ARCRJ). The Decima is a property record which has the name of the owner, the address and the assessed rental value for the property, for tax purposes. This exists in our database for the years 1849,1870,1878 and 1888, with 1878 not containing rental value as it was derived from a different source than the others. The Almanak is a city directory with individuals' names and, if they paid an additional fee, address and occupation. As a result, a lawyer or doctor would have their practice listed while a civil servant might just write the department they worked in. Our Almanak data exists for many years from 1845-1889, roughly every four or five years, with some larger gaps. The Voter rolls exist geographically by parish, but have rather discontinuous coverage, except for 1876, for which we have several parishes. This dataset contains a name, home address, occupation, and occasionally, the income and names of the voter's parents.

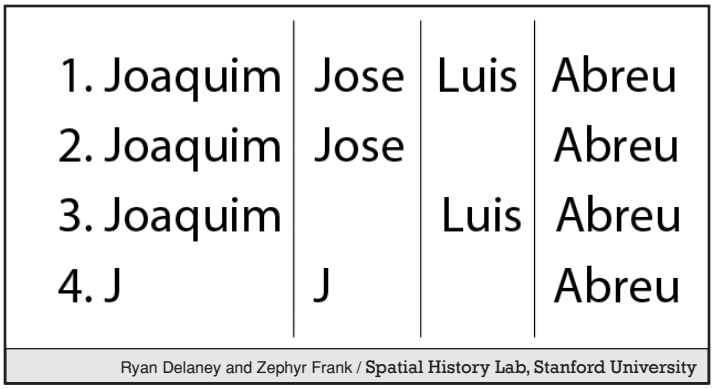

Figure 1 | Name Matcher. **The names above are potential variants that can be identified by the name-matching software.**

In order to help better interpret this data we have created a digital map of the city, which was primarily based upon an 1867 map of the city, in a geographic computer program called GIS, which in theory allows us to plot the addresses listed in each of our datasets, as long as our base map of the city contains accurate information of the streets' locations, names and the ranges of the buildings on the street. Additionally, we have developed a database to help us browse individuals, actual historical people may who exist within or between our multiple datasets. For example, a person may own multiple pieces of property in several years of our Decimas coverage, operate a business and so be listed in many of our Almanak entries, while also being a voter and appear as such. The trick is that wealthier Brazilians with lengthier names may in some sources write out their full name or appear as other permutations (see Figure 1)[1]. To overcome this we have developed a matching program that can handle these exceptions while weighting towards uncommon names (based on the fact that we have hundreds of thousands of names between the Almanak and Decimas), but also incorporating the number of name-words within an individual's name that match, and any additional information available. It is generally a proficient method, but has some shortcomings, as one might expect.

Name-Match Debugging

As the power of our project is derived from our ability to accumulate information between datasets - for example gathering the professions from the Almanak combined with the property

[1]Research Assistant, Stanford University Spatial History Lab

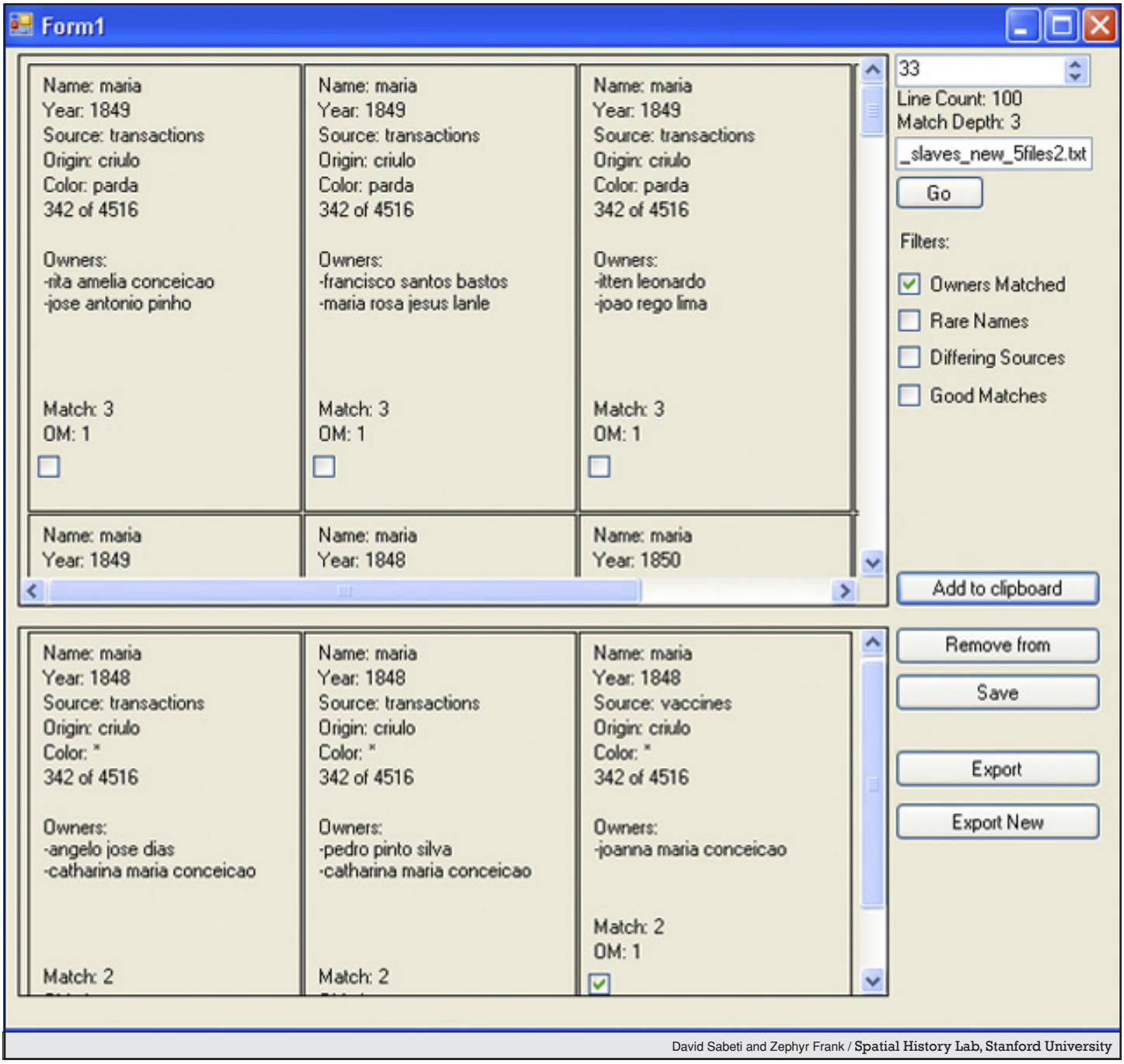

Figure 2 | Matched Slaves GUI.

values from the Decimas in order to determine which professional categories owned what amounts, and locations, of property - it is imperative for the project that we maintain some degree of error-checking over the matches generated based upon names. To this end, we developed a graphical utility capable of browsing the matches generated by our software, giving us the power to quickly inspect, and alter, them in a semi-automated fashion, while maintaining some manual control over the whole process. Figure 2 shows an adapted version of this program, which we have used to try and locate specific slaves within our multitude of datasets.

Slaves are far more difficult to identify because, as previously indicated, wealthier Brazilians tended to have longer names, so it follows that slaves, who on the other end of the spectrum, had only a first name. Because of this, the standards for a potential match, and the resultant quality, are far lower for slave matches and as such we have far less confidence in any automatically generated association. However, it does allow us to narrow down potential matches and let us manually review a manageable amount. As the name cannot be as heavily relied upon for slave matches, this adapted use allows us to highlight the other features of the debugger, including its ability to include different fields for comparison. For example in Figure 2, five datasets have been compared to look for

potential slave matches, resulting in the current display of three entries in the bottom panel that have been flagged as a possible match. The names are obviously identical, while the year, which refers to a date-of-birth for the slave, is also identical but we have the tolerance set to a couple of years, which is the best that can be historically expected for such information. Additionally, the origin and color of the slave, if given, are used to try and further increase the chance of a computer-generated match. Below color, the program indicates that 342 entries out of the 4507 in our five datasets are named maria, giving us a measure of the uniqueness of the slave's name.

In this particular case one of the owners in each of the first two entries, catharina maria conceicao, match perfectly and when we investigate it further we learn that one entry was a purchase and the other a sale; for a considerable profit. The third entry, from the vaccination records, has identical slave information and a very close owner name, and probably merits further investigation. While the only time we can truly be confident of a match is when the owners' names are identical, which occasionally happens, this at least provides a systematic, efficient, and rapid way of investigating possible matches of slaves.

Both in the role of approving matches of free citizens, and as a reviewer of possible slave matches, this program can easily rearrange and reorganize a computer's assessment of a match, based upon external historical information. It also allows us to use different quality settings in our database; perhaps for a large-scale aggregation we can use looser settings, while for a detailed small-scale question, we want only the highest-confidence matches – this program makes such variance not only possible, but efficient and quick.

We have also used this program to help identify where to conduct further archival research. Archival work is typically slow and time consuming, however, we used an index of names for which we knew archival material existed (without getting into too much detail, these archival files contained full inventories on individuals' property for execution of their wills), to find out which records would have the greatest chance of meshing with our other datasets, thus allowing us tell more complete stories. By matching the roughly 120 names in the index to our Almanak, Decimas, and Voter records, we found about twenty people who were in both the Almanak and Decimas, with seven of those also matching to the Voter records (we only had 2/7 parishes for that time period). With the list narrowed down to twenty good, and seven fantastic, matches we were able to maximize the use of our archival time, while at the same time increasing the odds of obtaining the best possible data for further research.

After the first wave of city-street generation was complete, the

Street-Debugging

dual process began of making the digital streets match the names in our datasets (for example Rua da Saude was standardized to r saude) and ensuring that the street ranges were accurate, so that a plotted address would actually be in its historical location. To achieve both of these ends, we developed a rapid error checking computer-tool that could interface with the commercial graphing software Tableau to help us quickly identify problematic areas, not only within the city, but within a given street. A key assumption of this tool is that we are more likely to be incorrect in how we digitally constructed a street than were the Brazilians recording their addresses for tax, or business, or voting, purposes. As a result, by identifying and quickly displaying the addresses that could not be plotted by the address-locator, we can see the problematic areas. Figure 3 is a screenshot of just such a visualization. The names of the streets, from the datasets, are displayed on the x-axis while the addresses on the street that were not plotted in the city appear on the y-axis. This single image allows us to quickly and efficiently adduce several key facts that would be nearly impossible to identify in an 80,000 element dataset. First, if a street has a solid bar, we know that both sides of the street are malfunctioning, while a dashed bar indicates that only one side of the street has a problem. Secondly, it helps us quickly find out which part of a street is not working; in Rua Nova Conde, the first street in the image, it is obvious that only the end of the street is failing, while the rest is working properly. If it appears that the entire street is missing, such as ld barroso, we can then quickly determine whether the problem is that the street is not spelled correctly (perhaps it should be ld barrosso, for example) or if we do not have the street in our city-system.

The speed with which this tool can be implemented (less than a second), and interpreted, is the key to its success; it has saved untold hours of frustration, and added a high-level of historical rigor to the later analysis, maps and visualizations which were produced further down the road.

The system works well enough that when a situation was uncovered that was not a simple typo, or an obvious range-error on our part, it was clear that there was a historical mystery to be uncovered. This is what happened when we tried to fix the seemingly innocuous problem of pc harmonia, which is the sixth street in Figure 3. While at first it may seem that one side near the end of that street is missing (similar to r nova conde), the plot thickens when you learn that Praca Harmonia is only a few buildings long, according to textual evidence and, especially, that it no longer existed at that time. Using this initial error-checking tool and visualization-method as a catalyst, we uncovered the history of Rua da Saude (Figure 4).

When our data begins, in 1845, portion H went by the name of Rua Sao Francisco de Prainha, and was numbered from right to left (East to West). From there began Rua Saude, which at that time ended at E, but was soon expanded to include F as well. Notably, G was undeveloped at this time and contained no addresses (more on this later).

In 1856 Rua Sao Francisco de Prainha and Rua Saude were combined to form a new Rua Saude and given new address numbers, so that Rua Sao Francisco de Prainha kept its numbers and Rua Saude's numbers progressed from there. The reason we can identify the year of this change as 1856 is because in 1878 the entire street-system of Rio de Janeiro was redone and some short historical facts about the streets were recorded in a book documenting primarily the 1878 changes. What was not indicated in this book however, was the location of the pre-1856 buildings. For example, we know where 84 Rua Saude is based on maps from 1867, but have no idea where addresses physically were before 1856. That is, until now. By using the aforementioned name-matched database, we were able to locate several Almanak entries

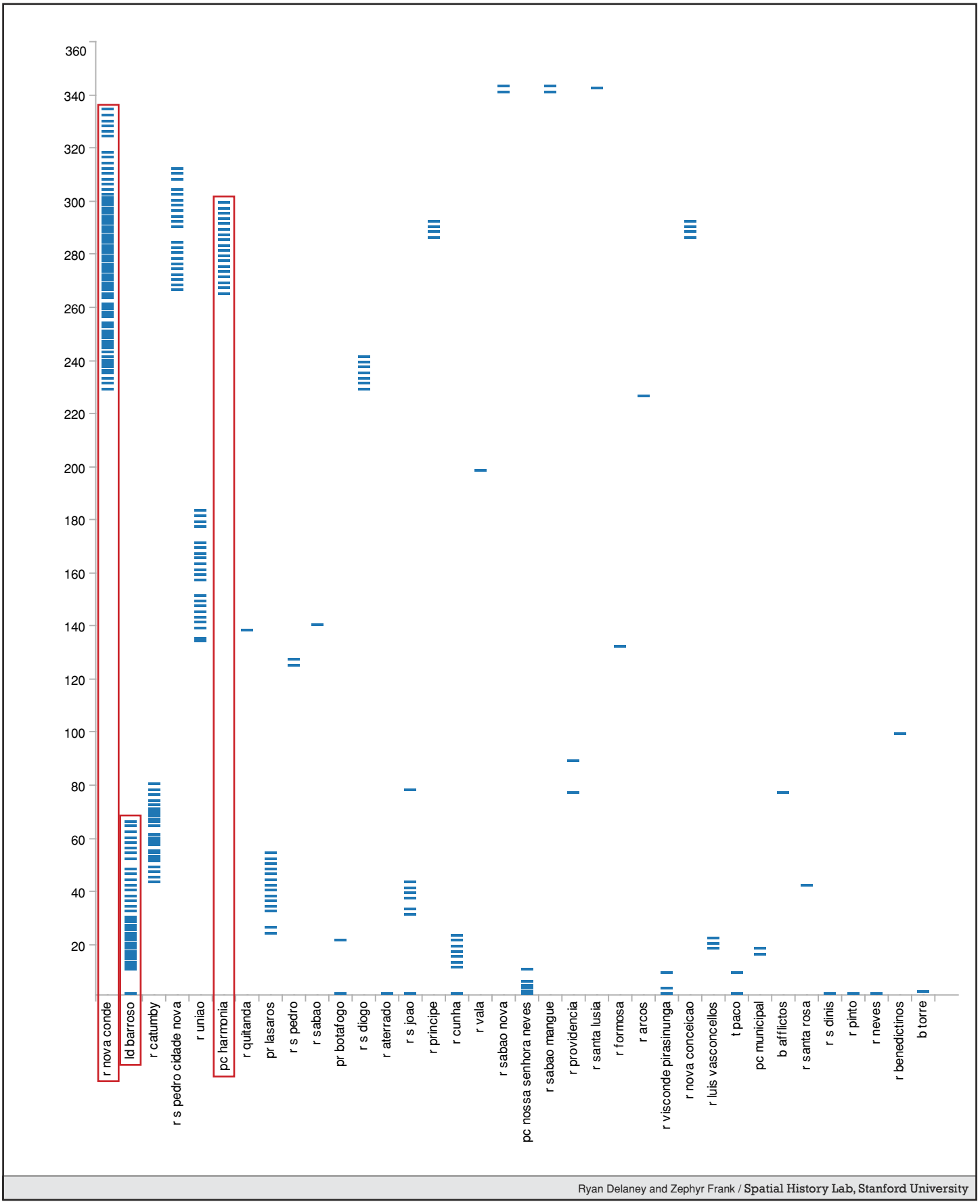

Figure 3 | Street Ranges.

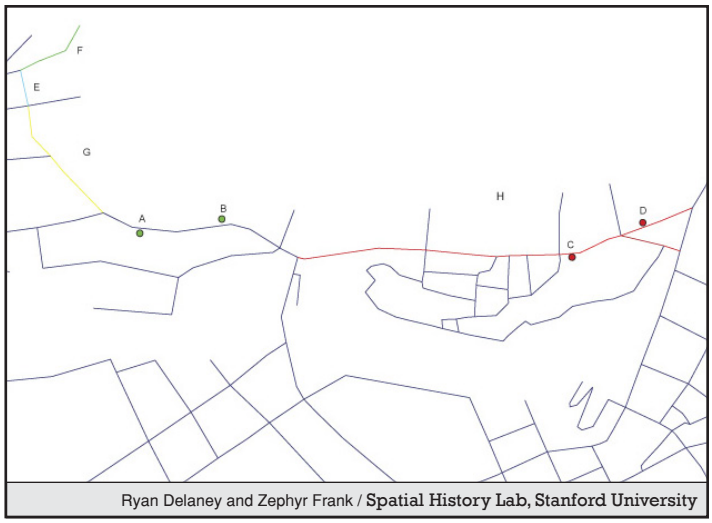

Figure 4 | Rua da Saude.

along Rua Saude in both 1854, before the change, and 1859, afterwards. Points A and B are examples of this match. In 1854 an identifiable individual listed their business at 61 Rua Saude, and in 1859 that same individual was located at 195 Rua Saude. On the other side of the street, point B demonstrates an individual who 'moved' from 10 Rua Saude to 84 Rua Saude. Points C and D show where the addresses would fall if they had in fact physically moved between these addresses, instead of merely reflecting the new numbering scheme.

Here is where the story gets complicated. Between 1856 and 1867 (the time of our first map), the area G was developed. We can tell this is the case because the numbering system of Rua Saude (post 1856) runs continuously through to section E, skipping cleanly over G. Perhaps in order to avoid renumbering the entire street again, section G was simply named as a separate street: Praca Harmonia, stuck right in the middle of another street. Praca Harmonia had its own number system, starting at one, and continued until the E portion of Rua Saude, where the numbers picked up seamlessly from before Praca Harmonia and the name once again continued as Rua Saude.

In 1874, section F was officially peeled-off from Rua Saude, and joined the street it more naturally seems to be part of, Rua Boa-Vista. 1874 is the official date, according to its history in the 1878 documentation, but it seems to, according to our datasets, have been already colloquially part of Rua Boa-Vista beforehand. At this time however, F still had its 1856 numbers, which ran from 303-359, and the rest of Rua Boa-Vista was numbered in the opposite direction, starting at Rua Saude with the number one.

Finally, in 1878, with the city-wide renumbering and naming enterprise, the situation was (mostly) resolved. Rua Saude was once again renumbered, this time incorporating Praca Harmonia, and ended at what was now Rua Boa-Vista. Section F became the new beginning of the Boa-Vista numbering scheme and that street was straightened-out. The wrinkle, and what started out the investigation into this street to begin with, was that, while accepting of the new numbering scheme, for years people still called the portion of Rua Saude that had been Praca Harmonia, by the latter name, which is the how high numbers for Praca Harmonia found their way into the street-debugger visualization in Figure 3.

Conclusions

Most visualizations represent the end point of a specific line of inquiry within a project; a map in a paper, an animation telling a previously determined story, or an interactive mechanism for browsing a project's findings. This, however, is only a small fraction of their usefulness, as they need not be restricted to presenting 'complicated' information to lay-people, but can be extremely useful in keeping the data rigorous and honest, while then helping to guide the research process by identifying new avenues for research. Visualizations and tools of this type, including those presented here, are far less flashy and complex, leading to their frequently being overlooked, but they can rapidly increase the productivity of any given task, literally from data entry to research direction.

1. Frank, Zephyr. Maria, Mother of Zenóbia: Onomastic Explorations of Slavery and Freedom in Rio de Janeiro, 1840s-1870s. Presented at the University of Michigan, 2008..

Supplementary Information is linked to the online version of the paper at http://www.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=11.

Acknowledgements I would like to thank the following people for their contributions: Zephyr Frank (Spatial History Project Associate Director and Principal Investigator on the Terrain of History project), Mithu Datta (Spatial History Lab GIS Specialist), Whitney Berry (Spatial History Lab GIS Research Assistant), and Spatial History Lab Research Assistants: Hannah Gilula, Lucas Manfield, and David Sabeti.

Author Information Correspondence and requests for materials should be addressed to Ryan Delaney (ryand@stanford.edu) or Zephyr Frank (zfrank@stanford.edu).