Reduction, approximation, and omission : preparing a dataset for visualisation

Autor(en): **Hornbake, Laure**

Objekttyp: **Article**

Zeitschrift: **Geschichte und Informatik = Histoire et informatique**

Band (Jahr): **18 (2015)**

PDF erstellt am: **27.05.2024**

Persistenter Link: https://doi.org/10.5169/seals-685433

REDUCTION, APPROXIMATION, AND OMISSION: PREPARING A DATASET FOR VISUALIZATION

Laura Hornbake

This paper demonstrates how creating digital visualizations for small historical datasets can lead to new interpretations that challenge existing periodizations and narratives. It takes as an example an animated map representing events over time and space. Describing the transformation of historical sources into a standardized dataset, it explores problems inherent in this process: the loss of detail through the reduction and standardization of data and the implications of this loss. It shows that sensitivity to these problems while creating a visualization can itself prompt new insights.

Introduction

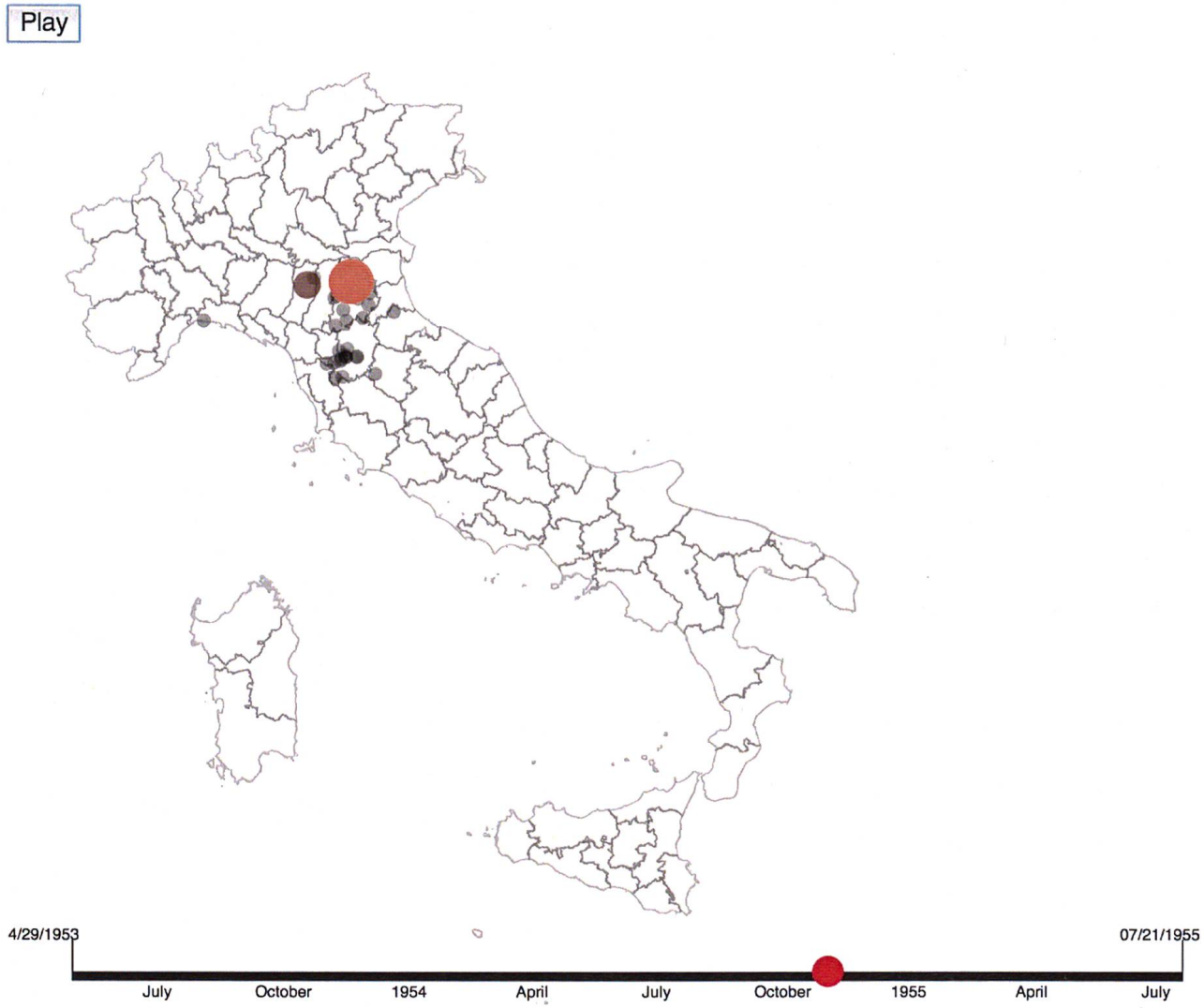
In the age of big data, what will be the fate of the kind of research that requires painstakingly looking through boxes of ephemera, extracting information from a tiny collection of sources, connecting the dots between varied, fragmentary sources? What can the digital visualization of data offer historians working not with large digital datasets, but with only a handful of sources? This paper presents a case study of a small-scale project in which fragments of information from primary sources are processed to create a standardized dataset and to generate a dynamic visualization, in order to examine the possibilities of these methods.

While not every research project is suited to generate digital visualizations, many historians would do well to consider what might be gained by preparing a small dataset and representing it visually. Visualization offers the possibility of making clear the complex patterns in chronology, in geography, in interactions between actors in a network, and in many other aspects of archival sources in ways that can challenge existing periodizations and narratives.[1] This paper champions small datasets and visualizations generated from them. It encourages historians to consider the conversion of primary sources into a dataset, examining both the promises and the pitfalls of such representations of information.

"State-ordered Evictions": A Case Study

"State-ordered Evictions" is an animated digital map that represents a campaign of evictions of associations in Cold War Italy, 1953–1955 (FIGURE 1).[2] It uses the JavaScript library D3.js to generate a Scalable Vector Graphic (SVG) of a base map of Italy and to graph data points that represent evictions onto that map at the coordinates of the events.[3] Moreover, the map uses animation to present the chronology of the events. Initially, the map shows only Italy and its internal provincial boundaries. When the animation is initiated with a button, the points become visible over a period of a few minutes. The timing of these appearances is scaled to the dates of the historical events. Thus the animation represents temporal-spatial data, displaying points that appear on the map in condensed animation time as the events unfolded in historical time.[4] It demonstrates that this campaign of evictions targeted specific central Italian provinces, and that these evictions were concentrated in a short-lived but intense period between the summers of 1954 and 1955.

FIGURE 1
"State-ordered Evictions of Associations, 1953–1955", pictured mid-animation.



This project to map evictions began while I was writing a dissertation on associations in central Italy, while working through a chapter on the importance of the experience of the 1950s in creating solidarity around embattled associations.[5] As I came across repeated references to evictions of associations, often described as a symptom of the "hard years" for the Italian left in the 1950s, I began to collect the details in a list. The list itself was unenlightening, a series of dates and descriptions that offered no clear insight, but that very obscurity prompted me to think about how I might analyze such information. Was there some pattern in these evictions that could help explain this campaign? Why were these evictions executed in central Italy in particular? How could I evaluate historical sources that explained the campaign as a politically-motivated strike at the communist heart of Italy? What kind of evidence could I muster to support or refute such claims? Were there correlations between distinctly political phenomena, for example, the electoral results of 1953 and the pattern of towns affected by evictions?

In presenting the campaign in my writing, it was not sufficiently compelling to simply sum up that dozens of evictions were executed in Emilia Romagna and Tuscany, yet I did not want to overwhelm readers with a deluge of place names and dates. Moreover, the places cited in my sources were small towns with names that would certainly be unfamiliar to most readers: Portile, Lastra a Signa, Baricella. Thus the first priority would be to represent the geographic distribution of events on a map of Italy, demonstrating that these place names corresponded to towns and villages in central Italy, predominantly in the provinces Bologna, Modena, Reggio Emilia, and Firenze. My hypothesis was that the evictions occurred in these places, which, I intended to explain in accompanying text, were also the areas of greatest support for the *Italian Communist Party* (PCI). From the initially modest ambition to illustrate my research with a map, grew a more fruitful project as I experimented further with preparing my original source material for analysis and with generating more complex digital visualizations. I discovered that the process of creating visualizations was both more problematic and more promising than I had originally anticipated.

This paper foregoes discussion of the technical aspects of generating the visualization to focus on the manipulation of data that must precede the creation of the visualization. The following sections explore the development of "State-ordered Evictions" and the problems of each phase, first detailing the stages of collecting data and processing it to create a dataset from primary sources, then discussing the further refinement of the data by testing the visualization and editing the dataset, and finally offering some conclusions.

Converting primary source material to a dataset

In order to perform any kind of analysis, the sources first had to be transformed into a standardized dataset, a process that began with assessing the source materials. I had uncovered a variety of contemporary newspaper articles, parliamentary papers, and secondary sources commenting on the evictions.[6] These sources ranged from lengthy descriptions of a single eviction, for example, extensive articles and photographs of the confrontation between police and protesters at the Casa del Popolo of Crevalcore in local and national newspapers, to passing notes of a town where evictions had also been enforced, without mention of a precise date or specific organization. While writing in narrative form on the subject, this heterogeneity created no problems: I could simply privilege the richer sources as examples, using the detail they offered to enrich my descriptions of the evictions and their aftermath while relegating the less specific sources to serve in statements about general trends, perhaps mentioning that evictions touched "dozens of other towns" in a province. This leeway with sources is a great advantage of textual over visual representation, particularly when working with archival sources. However, I hoped that by also analyzing my sources with the aid of a visualization, I might better understand how these evictions fit together not as isolated incidents, but as a pattern of similar events. In order to do that, I needed to reduce the varied details to a limited set of data fields. That implied suppressing the rich detail of some documents, but it also promised new insight that might be obscured by the particulars of any single event.

I had to determine the common features of the sources, the extent and limits of the data, and select an appropriate format that respected these characteristics and suited the desired type of visualization. Before launching into questions of the final data structure, I gathered up the various bits of information in an intermediate form, a spreadsheet (itself a form of visualization), where I could easily manipulate and revise fields, assess the data as it was entered, and export data in a variety of formats.[7] There I recorded the lists I had made about the evictions as I came across them in my research. For each of the more than seventy examples, I noted the place names that indicated the town in which an eviction occurred; the date it was executed; the organizations being evicted; any information about how the eviction was executed such as how many police officers were sent to enforce the order, whether citizens resisted the police, any resulting property damage or violence; the source of the information; and any additional notes. Fortunately these events were relatively simple and did not pose any complex questions such as date ranges or events involving multiple locations.[8]

However, as I noted above, this information was gleaned from scattered and extremely varied sources, and thus the quality and quantity of available details ranged widely. The authors of available sources did not always provide more than a place name; they often had simply compiled lists of the many towns where evictions had ousted voluntary associations from their seats in order to emphasize how widespread the phenomenon was. Yet others recorded painstaking detail about a single eviction. These asymmetries in the sources translated to asymmetry in the data. As a result, the spreadsheet of information gathered from these sources was full of blank fields where detailed information was absent. Because the fate of the dataset was to be input into a visualization that required a geographic location and a date, any points without at least these two properties had to be discarded. All other details became extraneous, recorded in a field of miscellaneous notes. It would remain attached to the data but without purpose, a vestigial trace of the original sources.

These processes of simplification and standardization of the data for the purposes of visualization do represent a loss.[9] They suppress the kinds of rich detail that many historians, myself included, use to create compelling narratives that describe how events unfold, not just where and when. In this case, these evictions were traumatic events for the communities that lost access to the rooms that were often their only public meeting spaces. To do justice to those experiences demands more than tallying the numbers of evictions, it requires extracting as much detail as possible from the few representative examples available. For this one facet of the historical analysis, however, we must throw out that detail, and hopefully come back to it with accompanying texts or other methods. I had already written at length about these events, and thus losing these details was not a concern.

Processing data

Once the information from primary sources was transformed into a standardized set of fields, I could begin to consider the way the data would function in the visualization and how the strings of text recorded in the spreadsheet might be interpreted.

The biggest challenge in processing the data for visualization was the transformation of place names into geographic coordinates that could be mapped. Fortunately, there are many services that can geocode addresses and place names, making the conversion of text strings to coordinates a simple matter of choosing an API, scripting a request to it and handling the response.[10] However, because the input data was cobbled together from a variety of types of sources, the process of making sense of the recorded place names could not be fully automated. It required some manual pre- and post-processing to disambiguate the names and control the quality of the output.

For example, authors of the historical sources I was using sometimes used colloquial, shortened versions of the name of a town which would have been understood only by their intended contemporary readers. A reference to the place "Ozzano" means little to anyone except those familiar with the small towns of the province of Bologna; while to the rest of the world the place is known as "Ozzano dell'Emilia". Moreover, it might not refer to Ozzano dell'Emilia at all, but to Ozzano Monferrato, a town in the northern region of Piemonte. This problem is likely to arise in other projects, which might be also complicated by using multilingual sources that use exonyms, by dealing with longer historical periods during which place names shift, or by using references to obsolete place names that have not been included in modern geographical databases. In order to automate the process of geocoding, the next step in producing mappable data, these ambiguities must be resolved. I found it necessary to return to the notes on my sources, to read for contextual cues to identify several ambiguous names. Fortunately most of my sources were specific to an Italian province. Once the additional field of province was added to the data, I could search for an unequivocal name and province combination, returning a single latitude and longitude coordinate pair.

However, by geocoding the place name, indicating a town or village where an eviction occurred, I was approximating the location of events. These evictions actually occurred at a specific building, an exact street address within the towns noted. However, that information was unavailable for all but a handful of items. This could have been a more serious problem for mapping on a smaller scale, for example, a map with an extent of one or a few towns. In such a case, imprecision in location could introduce significant uncertainty in the results. However, on a map of all of Italy,

using a town as an approximate location for an event produces results indistinguishable from more precise coordinates. The decision about whether a proposed approximation will significantly alter the results of an analysis, must be done on a case by case basis. For the purposes of this project, a town was a reasonable approximation of location, allowing me to append an approximated geographic coordinate pair to each event.

With all the data fields completed, the dataset could finally be exported in a format that functioned with the desired visualization tools. I chose to export the data as a GeoJSON file, and created a script to convert my spreadsheet into well-formatted GeoJSON.[11] This allowed me to use this dataset with a base map in ESRI shapefile format, a geospatial vector data format, using Geographic Information System (GIS) software.

Test Visualization and Refinement

With the dataset prepared in a format appropriate for visualization, I could then check the results with a test visualization and re-process data in as many iterations as were needed to achieve satisfactory results. This step should make any glaring errors clear, for example, if the process of geocoding has not produced acceptable results, points might appear in unexpected places. More significantly for the purposes of analysis, the first visualization of the dataset should clarify the general trends in the data and identify any outliers, points that are significantly distant from the other data. In the case of the "State-ordered evictions" map, the test visualization revealed points that were significantly distant from the others in both of the two variables represented, time and space. While most of the events represented on the map of Italy clustered around cities in Emilia Romagna and Tuscany, a handful were in other regions. This deviation in the geographic distribution was of little concern: a few exceptions could not diminish the clear trend, that the evictions were overwhelmingly located in central Italy. Given this geospatial trend, I might have considered omitting the points outside Emilia Romagna and Tuscany and changing the extent of the map to zoom in on only central Italy. However, by using a map of all of Italy, the argument of the visualization is decidedly different from that of a detailed area. My thesis is that these events represent national political maneuvers and illustrate something of fundamental importance about the relationship between central Italian political subcultures to national Italian politics. Thus the map answers not the question in which towns in central Italy were evictions enforced, but what parts of Italy bore the brunt of a particular national political agenda.

Considering the temporal data prompted similar questions of what to include or omit, which would frame the interpretation of the dataset.

The animation of the points appearing on the map showed a few initial events, followed by a long pause of no events, then an explosion of most of the events in a burst, followed by another extended pause and finally one single event. The long pauses indicated the temporal distance of the first few and last one of the events from the vast majority of the data, suggesting that the leading and trailing outliers should be more closely examined.

The question of how to handle the late outlier was an easy matter. The notes attached to the data point indicated that the eviction was not executed at all, but postponed. Its inclusion in the dataset was, in fact, an error, and it could be omitted without concern. The early group of outliers was more puzzling. What did these deviations from the general pattern mean? Were they relevant? To decide how to handle them, I first considered the assumptions I had made in researching and preparing the dataset: were there lacunae in the research that might have produced gaps in the data? While my research was thorough, I could never claim any certainty that it represented the complete record of all evictions. Were the sources for the outlying data different in some way from the sources for the other data? Were there problems in the way I conceived of the category "eviction"? Perhaps not all of these events belonged together as examples of a single phenomenon. Were there other ways of grouping this data? Was further research likely to uncover differences between these outliers and the rest of the data? As I formulated and considered these questions, I contemplated the periodization of the project. A possible explanation for the long pause in evictions between July 1953 and June 1954 might be found in national politics: the successive reshuffling of governments that removed Mario Scelba, architect of the strategy of evictions, from the office of Minister of the Interior in July 1953 for seven months, and then his resumption of that office when he also became Prime Minister (February 10, 1954 – July 6, 1955). In support of this interpretation there is the highly suggestive coincidence of the last date of evictions among the early group, which is the same as the last day of Scelba's leadership at the Ministry of the Interior. This interpretation suggests that the early points were not outliers at all, but important indicators of Scelba's role in pushing the campaign of evictions. This correlation was not clear, however, from the test visualization itself, which lacked labels or other links to this political periodization.

In order to improve the visualization, I could remove the outliers from the dataset, thereby avoiding the pauses of inactivity that might lose readers' attention or be misread as a technical error; or I might incorporate more information to link the periods of activity and inactivity to political tenures. Such decisions will be different for every project and

every dataset, but a good guideline might be simply to be scrupulous: to not use omission of data points to silence alternative interpretations of the information. In this case, I initially chose to omit early and late outliers, to refocus the project on 1954–1955 to make the animated visualization proceed smoothly. However, I reconsidered this decision while writing the paragraphs above, concerned that I had privileged the neatness of the visualization over the historical arguments that had originally inspired the creation of the visualization. Another of the great benefits of these methods is that they are flexible, allowing such experimentation and revision. This experimentation prompted me to consider the data in new ways, generating unexpected and important observations.[12]

Conclusions

The phases described above have brought up several important concerns for historians who wish to prepare a data visualization from their research. The theme that ties together these concerns is loss: the loss of details in reducing sources to limited data fields, the loss of accuracy in accepting approximate locations, the loss of exemplars in trimming outliers from the dataset. Such sensitivity to what any particular method requires us to give up is well-founded, though it should not become an obstacle to experimentation with new methods. As this case study demonstrates, information was not as much lost as it was traded for different kinds of insights that writing about the integral sources did not reveal.

In fact, one of the promises of data visualization for historical research is that it can be inclusive: datasets can incorporate information from more sources than one could possibly reference in a purely textual form. While writing about the period of evictions, I wrote at length about what I viewed as the best sources, those that offered the most vivid detail, the most interesting phrases to cite, while relegating the less well-documented cases to a passing mention or footnote. Those lesser examples stand on equal footing in the data visualization. In this sense, the dataset is in fact richer than the textual presentation of the same sources.

Moreover, as this paper has demonstrated, the promise of data visualization is that it makes visible obscured patterns. Preparing data for visualization prompted me to significantly revise my interpretations. In my attempts to understand the gap between an initial early cluster and the remainder of events, I discovered a correlation between the political tenure of Christian Democrat Mario Scelba as Minister of the Interior and the campaigns of evictions. This highlights the influence of Scelba in shaping policies that pitted the national government against the local forces that represented its most defiant opposition. It demands that we

look more closely at struggles between local and national authorities and at the role of personal antagonisms which may challenge interpretations of Cold War politics. The potential to generate such insights makes visualizations far more valuable than mere illustrations.

While these methods are not lacking difficulties, I believe that even with small amounts of data, the trade-off between integrity of historical sources and the analytic potential of a reduced dataset may often prove valuable. I would encourage other historians to undertake the processes of data entry and processing for small datasets, thereby enriching both their own research projects with new interpretations of their sources as well as securing a place in field of digital history for smaller studies.

1 There are by now many great examples of the use of visualizations to reveal patterns in large datasets. See for example: "Mapping the Republic of Letters" project at Stanford University's Center for Spatial and Textual Analysis (CESTA), <http://republicofletters.stanford.edu>, or the "Mapping Texts" project produced in cooperation between Stanford University and the University of North Texas, <http://mappingtexts.org>.

2 This article refers throughout to version 1.3, which will remain available at: <http://laura.hornbake.com/projects/cultureWars/map1,3> (last visited 2/10/2014). The latest deployment of the project, which will continue to be revised, is: <http://laura.hornbake.com/map> (last visited 2/10/2014).

3 Michael Bostock, Vadim Ogievetsky, Jeffrey Heer, D3: Data-Driven Documents, in: IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis), 2011, <http://vis.stanford.edu/papers/d3> (last visited 2/10/2014).

4 For critical view of the limitations of these types of representations see Ian N. Gregory, A map is just a bad graph. Why spatial statistics are important in historical GIS, in: Anne Kelly Knowles (ed.), Placing history. How maps, spatial data and GIS are changing historical scholarship, Redlands, California 2008, pp. 123–149.

5 Laura Jeanne Hornbake, Community, place, and cultural battles. Associational life in central Italy, 1945-1968. New York 2013, <http://hdl.handle.net/10022/AC:P:21627> (last visited 2/10/2014).

6 Sources such as Luigi Arbizzani's contemporary article on the phenomenon, see: Luigi Arbizzani, Lunga vita alle case del popolo, in: Emilia, Bologna 1955; and Nilde Jotti, I lavoratori in difesa della Case del Popolo, Reggio Emilia 1955; provided the lists of evictions in Emilia Romagna that inspired further research into the topic. To uncover more details on those reported evictions and to obtain a more complete list of similar events elsewhere in Italy, I searched the digitized archives of parliamentary records: Camera dei deputati, Lavori Parlamentari, and Camera dei deputati, Atti e Documenti: <http://storia.camera.it> (last visited 2/10/2014); the digital archives of the national newspapers L'Unità <http://archivio.unita.it> (last visited 2/10/2014), and La Stampa <http://www.lastampa.it/archivio-storico/> (last visited 2/10/2014); and researched secondary sources on associationism throughout Italy. Further research on the incidents that were the results of these searches included local periodicals and archives, when available.

7 While more complex projects involving relations between various fields and data points might require careful structuring of data in databases, a spreadsheet is a sufficient tool for processing non-relational data.

8 On the problems of representing historical time, see Manfred Thaller, Which? What? When? On the virtual representation of time, in: Mark Greengrass (ed.), The virtual representation of the past, Surrey 2008, pp. 115–124. For a more sophisticated approach to time, see the reports of the "Temporal Modeling Project" at the University of Virginia, <http://www2.iath.virginia.edu/time/reports/index.html> (last visited 2/10/2014).

9 Lev Manovich discusses reduction in information visualization and proposes methods that do not require reduction in: Lev Manovich, What Is Visualization?, in: Visual Studies, 26(1), 2011, pp. 36–49.

10 I experimented with both the GeoNames geographical database API and the Google Geocoding API.

11 I published this tool as a web application to make it available to anyone who wishes to replicate this process of mapping. See Laura Hornbake, GSS to GeoJSON, <http://laura.hornbake.com/projects/Tools/gssToGeojson> (last visited 2/10/2014).

12 William G. Thomas argues that the unexpected, the outcome of speculative assays is at the heart of digital humanities work. He suggests, paraphrasing Jerome McGann that, "...if you have produced what you thought you would, perhaps you've not created anything really..." in: William G. Thomas, What we think we will build and what we build in Digital Humanities, in: Journal of Digital Humanities, 2012, <http://journalofdigitalhumanities.org/1-1/what-we-think-we-will-build-and-what-we-build-in-digital-humanities-by-will-thomas/> (last visited 2/10/2014).