Teaching History in the Digital Age
Kelly, T. Mills

Published by University of Michigan Press

Kelly, T. Mills.
Teaching History in the Digital Age.

For additional information about this book
https://muse.jhu.edu/book/22708

For content related to this chapter
https://muse.jhu.edu/related_content?type=book&id=817458
Analyzing

Making Sense of a Million Sources

When we think about the future of historical research in the age of the huge digital libraries that are currently under construction, we will face with what I sometimes think of as the Klofáč-Kramář dilemma. In the late nineteenth and early twentieth centuries, Václav Klofáč and Karel Kramář were prominent Czech politicians—first in the Austro-Hungarian Empire, and later in Czechoslovakia. Because neither man became president or had much of a reputation outside of parochial Czech political circles, you should not feel guilty if you have never heard of either one. But for historians of modern Czech politics (like me), they are central figures in the historical narrative of the first Czechoslovak republic. Kramář is fairly easy to research using conventional methods. The Kramář collection at the Archive of the National Museum (Archiv Národního muzea) in Prague contains more than one hundred boxes of manuscript sources from his life and career, and several other major archives in Prague, Brno, and Vienna contain significant numbers of primary sources devoted to Kramář. These primary sources are not (yet) digitized, and so one must journey to central Europe to see them, but they are reasonably well organized and readily available to researchers. There are several biographies of Kramář, at least a couple of which have real scholarly merit, and at last count scholars have published dozens of articles on Kramář. Historians know how to work with a subject such as Kramář and how to train our students how to work in archival collections like those devoted to his life.

We also know how to deal with a subject like Klofáč, even though he
is much more difficult to pin down in the archives. Like Kramář, Klofáč has been the subject of several biographies; numerous (though fewer) scholarly articles; and, like his competitor for the attentions of Czech voters, he shows up regularly in histories of Czech politics from the 1880s to the beginning of the Second World War. However, researching Klofáč is a more difficult archival problem. Unlike Kramář, there is no major collection of Klofáč documents for the simple reason that when the Germans took full control of Czechoslovakia in the spring of 1939, Klofáč burned the vast majority of his personal archive, and after the Communist take-over of Czechoslovakia, his son destroyed the rest (his father having died during the war). The intent of these destructive acts was to keep these documents—many of which might have been used to implicate friends and colleagues—out of the hands of agents of repressive regimes. Thus, there are no shelves groaning under the weight of hundreds of boxes of Klofáč sources. But this lack of available sources does not mean Klofáč is invisible to historians—just more difficult to come to grips with. Researching Klofáč is much more of a scavenger hunt with many more miles of travel involved, but he lives in the collections of many dispersed archives around central Europe in letters he mailed to others, in articles he wrote for newspapers, in the minutes of meetings of the political party that he led for two decades, and in the reports of Austrian government spies who tracked him from his appearance on the political stage until his arrest for treason in 1914. Then there are the extensive transcripts of his trial for treason which include a lot of detailed testimony about his life and political activities. The “Klofáč archive” that is dispersed across all these repositories has also not been digitized, but almost certainly will be one day, along with the more easily accessed Kramář materials.

When that happens, as it almost surely will, given that these two men were founding fathers of the modern Czech state, what will historians do with those thousands and possibly tens of thousands of primary sources? At one level, access to and use of the Kramář archive will only be opened up and sped up. Instead of traveling to Prague, one will be able to work with the Kramář materials at a distance, and will be able to search through that archive with more speed and efficiency. By contrast, access to the Klofáč archive will be opened and sped up, but also historians will be able to create something like a unified collection through the aggregation of sources from those now scattered across central Europe. At this basic level, use of material from the lives of both Klofáč and Kramář will be easier for
Analyzing historians. What will be different is that if these materials are all marked up properly when they are digitized, it will be possible for historians to do much more than access these materials faster and from the comforts of home. We will also be able to start triangulating across a wide range of archival repositories that we had not previously thought of. So, for instance, was Kramář or Klofáč mentioned in a document in a collection we did not know existed? And if so, why and in what context? Similarly, we could chart the ebb and flow of a public figure’s level of activity and/or interest value (to the public, to the secret police, in the media) by tracking how often he or she shows up in the sources.

Already, data-mining software makes it possible to link sources on the basis of date, location, names, institutional affiliations, and all the other ways historians triangulate between and among sources. In the past we have had to do that triangulation by hand, and so making these connections is often a laborious and imperfect process. With each passing week, more and more historical data appears online marked up in ways that make it possible for us to use new software tools to work with these data. Now software can make connections for us and possibly even propose new ways of thinking about things such as relationships between individuals. As Greg Crane pointed out several years ago, “Already the books in a digital library are beginning to read one another and to confer among themselves before creating a new synthetic document for review by their human readers.” While Crane was writing about books “reading” one another, the same can already be said for non-book sources as well. The resulting “recombinant documents,” as Crane calls them, offer the historian very different ways to look at and think about historical sources. Do we know what to do with such recombinant documents? And do we know how to train students how to work with such an overwhelming corpus of sources?

The answer, I propose, is both yes and no at the same time.

The “yes” part of the answer is that today’s historians are well versed in thinking critically about historical sources, and those skills are not made obsolete by recombinant sources or by historical information presented to us in other ways such as can be done with sophisticated visualization software. But as useful as our current skills are, they are predicated on the form of the primary source. As Sam Wineburg has already demonstrated in his research on how historians think, historians approach primary sources in certain discipline-specific ways. Watch any historian read a letter written 100 years ago and you will see her or him check first for contextual data
such as the date the letter was written, the author's name, the recipient's name, the place the letter was written, where it was mailed from and to, and any other data such as an institutional letterhead on the paper that might be available. Only when all of these bits of information have been mined from the source will the expert learner/historian begin reading the body of the letter. Novice learners, by contrast, tend to launch right into an examination of the main body of the source, coming back to contextual data only later, if at all. Our goal in teaching historical methods is to train students to learn the same skills we have developed over many years of study and, we hope, to turn that learning into reading and analytical strategies that are as reflexive as ours are.

For example, although a personal letter increasingly seems like an artifact from a bygone era to students, they still know what a letter is, and so they apply whatever skills they have learned, just as we do, to the source in its original form. As any history teacher knows, students typically want the analysis of a personal letter to be relatively simple and straightforward. They want to know who the letter was from, whom it was addressed to, why it was written, and what it says. If some of the content is inflammatory or salacious, they (and we) naturally gravitate to that aspect of the letter. But if the letter seems pretty mundane on a first reading, they may quickly decide, “not much to learn here” and move on to the next source. Teaching them to read more carefully, to mine useful information from the seemingly mundane, is more difficult. Digital media make it possible to construct simple exercises that introduce students to the idea that something as seemingly simple and straightforward as a letter or a short telegram can be quite complex when read carefully.

Several years ago, back in the Web 1.0 era, my colleague Kelly Schrum and I designed a series of what at the time seemed like very interactive online exercises for students to introduce them to the complexities of working with historical documents such as personal letters, newspapers, maps, and so on. For the exercise on reading personal letters I selected a brief letter sent from Prague in the spring of 1939, just after the German takeover of the rump Czech state that had survived the Munich Conference disaster. This letter, sent from an American student to his cousin back in the United States, merrily recounts his bicycle ride across the Czech-German border, through the hills of northern Bohemia, and into Prague. Once in Prague he was witness to the German troops and tanks riding into town, and the
inauguration of the Nazi Protectorate government. The letter is chatty, and the author breezily recounts the difficulties of his travel, some caused by the German takeover, some caused by the steep hills. I have assigned this letter to my students many times and only rarely do they extract any worthwhile insights from the text on the first reading. In our online exercise, my colleague and I posted up two versions of the letter—one without commentary, then a second with commentary from a historian (me). As the user drags his or her cursor over the text, the historian’s commentary appears. For example, in the second paragraph, the author writes: “At first the Czechs got sore, blocked the streets, shook their fists at the troops, sang their national anthem, but when they saw more and more German troops pouring in, they saw their cause was hopeless and went back to their work.” The text from the historian that pops up offers this commentary: “Historians are very interested in the supposed lack of resistance to the Nazis by Czech citizens. Kistler’s account provides some verification of the common view that most Czechs simply did not resist. This is not news to specialists, but does provide further validation of one version of what happened in Prague.” This simple use of technology to give students a glimpse of more analytical reading strategies often prompts them to be much more analytical with later sources I give them. For example, when I next give them a personal letter to read and analyze, they are much more likely to think carefully about the historical context within which the letter was produced, asking themselves questions such as how the events swirling around the author might have (or might not have) colored his perceptions of what he was seeing, and why.

We are, as Greg Crane points out, on the cusp of an entirely more complex set of possibilities when it comes to using digital media to create teaching and learning opportunities for students than the one just described. For more than a decade, advocates of hypertext have promoted its value as a catalyst of new forms of reading. But hypertext built in HTML is inherently limited in its ability to create new forms of historical presentation, because HTML is a presentation language that describes what something will look like online (and what it is connected to elsewhere on the Internet). By contrast, XML describes the content it is marking up. A simple example can demonstrate the difference between the two languages. If one were to create a page of famous speeches by American presidents, the beginning of the source code for the page might look like the code below.
<html>
<head><title>Presidential Speeches</title></head>
<body>
<h2>Presidential Speeches</h2>
<hr>
<em>Farewell Address</em>, <b>George Washington</b>, 1796
<em>Gettysburg Address</em>, <b>Abraham Lincoln</b>, 1863
<em>Declaration of War</em>, <b>Franklin D. Roosevelt</b>, 1941
</body>
</html>

But in XML, the coding of that same content might look like the following.

<speeches>
  <speech>
    <title>Farewell Address</title>
    <author>George Washington</author>
    <year>1796</year>
  </speech>

  <speech>
    <title>Gettysburg Address</title>
    <author>Abraham Lincoln</author>
    <year>1863</year>
  </speech>

  <speech>
    <title>Declaration of War</title>
    <author>Franklin D. Roosevelt</author>
    <year>1941</year>
  </speech>
</speeches>

In the HTML code, the titles of the speeches are rendered in italics, while the names of the presidents are rendered in bold face type. In the XML
code, none of this formatting exists, because such formatting decisions can be made by the user elsewhere, using cascading style sheets or other forms of formatting that, for instance, might define all titles of speeches as rendering in bold face, while the authors’ names appear in italics. However, much more useful than formatting of text is the ability of users to extract information from historical documents marked up in XML by such fields as author, title, or year. In other words, in XML, the content and the form that content appears in are two separate things. XML makes recombinant documents possible.

To be sure, users of hypertextual documents—whether created in HTML or XML—do read in different ways than those whose documents contain no links. Whenever we click on a link we bounce from one source to another, sometimes returning to the original, sometimes not, but the various sources we see on-screen retain their shape and form—it is just how we get there and away from there that changes. But what happens when, instead of jumping from one document to another along hypertext links, our screen displays a recombinant document that has been parsed in ways that show only its references to a particular event—say a meeting of the party leadership to decide whether or not to form an electoral alliance with a rival party—and is simply part of a list of such references along with chunks of text that are devoted to that meeting? In other words, on-screen one might find the sentence from a letter written by a party leader dismissing the meeting as worthless, a three-sentence assessment of the meeting written by another participant who did not attend, but heard about it from someone else, and a half dozen other bits and pieces drawn from the archives of other politicians (and perhaps also government spies). Also on-screen might be a map showing not only the location of the meeting, but also the locations where each letter originated and/or was delivered, a time line showing the time sequence for each letter, and a link to the minutes of the meeting from the party’s archive. What happens to the reflexive skill we have developed for reading letters when the letter is no longer a letter, but has been reduced to chunks and bits of a recombinant document?

This description of what one might find on-screen is not a fantasy of the future—it is already doable. The only impediment to the display of such recombinant documents is the marking up of the relevant documents, the writing of algorithms to scrape the relevant information from those documents, and a user interface that displays the scraped information in a way that is easy to read and work with. Such scraping algorithms already ex-
ist. Perhaps the most popular is the search engine Google. Think for a minute about what you see after typing some keywords into the Google search bar. The screen shows a combination of highlighted text, a snippet of information from the web resource (document, website, discussion forum posting, etc.), and some other relevant metadata including the current URL of the resource. Google also gives you the option of viewing your search results on a time line, or in other ways such as the Wonder Wheel. Now, instead of the search results you see from a Google search, imagine that you are working on a research project on the history of slavery in America and you decide to search through the personal correspondence of Thomas Jefferson for references to slaves and slavery. Instead of looking at search returns that take you to each individual letter Jefferson wrote in which he mentions slaves or slavery, you get a document that provides you with the paragraphs from those letters where slaves or slavery are discussed along with relevant contextual data such as dates, locations, recipient information, and so on. These paragraphs could be arranged in a variety of ways—chronologically, as part of a series of back and forth exchanges with individuals, or any other way you might choose. They would include a link back to the full document so that you could read the full text of whichever letter you chose, and you could view the sources either along a time line, on a map, or just as chunks of text on-screen. Already Google’s book search makes a very limited version of such a recombinant document possible. Using Google Book Search you can search through a book for chunks of text that contain a word or phrase. A search on “slaves” in the 1829 text of Jefferson’s Notes on Virginia turns up fifteen pages where the word “slaves” appears. What Google Book Search will not do is allow you to search across multiple texts simultaneously. Once librarians, archivists, historians, and the general public have completed the task of marking up corpuses of text like Jefferson’s correspondence, it will be relatively easy to produce recombinant sources that work as described earlier. This is the world we need to train our students for, but first we ourselves need to learn how to make use of these tools, and we need to be part of the discussion of how they are implemented in our field.

More historical texts than can easily be counted have already been scanned and placed online: Google has scanned more than 20 million books, and is scanning new works at the rate of 1,000 pages per hour; other smaller projects such as the Open Content Alliance, the Million Books Project, and others are likewise making millions of books available online;
digital repositories of scholarly articles are also growing at a rate almost unimaginable just a few years ago.\textsuperscript{9} For example, as of May 11, 2010, the JSTOR database contained 37,307,998 pages from 6,219,336 articles in 1,239 journals.\textsuperscript{10} LexisNexis claims to offer access to “billions of searchable documents and records.”\textsuperscript{11} The number of digitized primary sources is growing at a similarly rapid rate. The Europeana.eu project aggregates more than 20 million digitized primary sources.\textsuperscript{12} The American Memory Project at the Library of Congress now offers more than 15 million primary sources in digital form, and just one newspaper scanning project—ProQuest Historical Newspapers—offers access to more than 25 million digitized pages.\textsuperscript{13} These numbers do not even take into account the amount of digital data we are producing every year that future historians will have to grapple with—perhaps as many as 1,200 exabytes in 2010 with growth rates as high as 60 percent per year predicted through 2014.\textsuperscript{14} One can only hazard a guess at how many historical primary sources will be available in digital form a decade from now when today’s undergraduate students will be writing their dissertations, teaching high school history classes, creating museum exhibits, or building their own digital exhibitions just for fun.

Recombinant sources such as those described earlier are just one way that historians and history students are and will be working with digitized sources in the coming decade. As of this writing, there are not enough historical sources marked up in XML format nor are the analytical algorithms up to the tasks we would like to set for them, but it is only a matter of time—probably just a few years—before Crane’s vision can be realized, and students need to be ready. In the rest of this chapter I want to describe a few data- and text-mining methods that can be used right now to begin to make sense of the digitalized information already available.

\textit{Geographic Interfaces}

Perhaps the most common lament of the history teacher—after complaints about their students’ writing—are complaints about how little our students know of or understand about geography, especially historical geography. It might be fashionable to blame GPS devices for turning us into a society of geographic illiterates who cannot read a map to save our lives, preferring instead to just follow the soothing voice of GPS devices, but concerns about student geographic illiteracy did not begin with the appearance of inexpensive directional aids on the market several years ago.\textsuperscript{15}
Historians have been worrying about this problem for decades, if not longer.\textsuperscript{16} At the most basic level we want students to be able to read a map; to decode some, if not all, of the information it contains; and to understand that a map is a historical source that makes an argument all its own.\textsuperscript{17} At a more sophisticated level, we want students to understand that human actions have been constrained or abetted by geographic realities. It is this latter goal that can best be served through digital tools because those tools allow us to create visual representations of historical information that are explicitly linked to geography. These geographic visualization systems can be either quite simple such as creating a layer for Google Earth, or quite complex, using the most sophisticated geographic information processing systems such as those provided by companies like Esri. These latter systems also make it possible to mine large databases of geotagged information to create sophisticated maps of data. Among the simplest examples of how geographic visualizations can be used to help students make sense of events in the past are simple layers created for Google Earth or other similar mapping interfaces. The community of users creating historical layers for Google Earth is quite large, and the number of new layers produced each day continues to grow at a rapid pace.

Students working on a particular research project can be well served by examining the layers available on their particular topic. For instance, a student researching the U2 incident during the Cold War might find his or her way to a Google Earth layer that maps out the diary entries of Frances Gary Powers throughout his career (with a particular emphasis on the period 1958–1962), and includes photographs from Powers’s personal collection, as well as approximate route maps for his flights over the Soviet Union.\textsuperscript{18} Seeing the events on the globe while reading the diary entries can help our student to understand how the surveillance program had to take into account the great distances involved in overflying the Soviet Union. Because Power’s diary entries are all geolocated, the student researcher can also see his career as a pilot in geographic space, not merely as words on a page. Many, if not most, of the historical layers created for Google Earth exist because amateur historians create them. As a result, encouraging students to use these layers as historical sources without some sort of training in how to pay close attention to what they find there is akin to the problems of turning them loose on search engines discussed in chapter 2. Among the questions they ought to be asking of this particular source include: are the Powers diary entries provided in this layer complete or edited; have all of the
diary entries been added to the map layer, or only those that make a point
the creator of this map layer wants to make; and who created this resource?

A more sophisticated version of this same sort of project is Light and
Shadows: Emma Goldman 1910–1916.19 This blending of geography and
historical sources provides users with a map of the United States with a
pin in the map indicating “all the places in America that the anarchist
Emma Goldman gave talks, and all the topics she spoke on” between 1910
and 1916. A temporal slider across the bottom of the screen allows users
to limit the number of pins in the map to a particular year or even part
year. For instance, if the user selects just the year 1910, thirty-six pins are
displayed: each of them offering information about one or more of the
events on Goldman’s schedule. Wherever possible, the project team has
embedded links to documents from the Goldman archive, including news-
paper stories and texts of lectures given by Goldman. One glance at the
map indicates how well travelled Goldman was as a speaker. During 1910
alone, she spoke (or attempted to speak) up and down the West Coast,
through the mountain West, across the upper plains, down into Iowa and
Missouri, up along the Great Lakes, throughout the Northeast, and down
the mid-Atlantic coast as far as Washington, D.C. A student looking at
this map might well ask how someone could cover so much territory in
the United States in 1910, what forms of transportation she might have
used (train, horse, boat), who paid for all that travel, and why Goldman
spoke in certain locations and not others? By drilling down even further,
one can see that the bulk of Goldman’s activities in 1910 took place before
the end of June, with almost no speaking engagements in the second half
of the year. This finding makes it even more surprising that she could
have covered so much territory in just six months, and begs the question
of why her speaking trailed off in the second half of the year? The answer
to this question is found on the website by clicking on one of the pins for
New York City—Goldman was in the hospital (under an assumed name)
in the summer of 1910, recovering from a broken kneecap. Or one could
look at the site in a different way, focusing on only one location—for ex-
ample, St. Louis—to see that Goldman spoke or attempted to speak there
on twenty-one different occasions, for which the project offers more than
thirty related primary sources. Among the lessons students can learn from
working with this geographic interface are that transportation was perhaps
more efficient than they might have imagined it to be 100 years ago, that
politicians often travel to locations where they have willing audiences, that
anarchist sentiment seems to have been spread across the United States just after the turn of the twentieth century, and that this sentiment seemed to be clustered in industrial centers.

At a more sophisticated level, historians and geographers have created web-accessible interfaces that allow users to examine large datasets in geographic space. For example, the project NS-Crimes in Vienna offers users the ability to examine a large database on the expulsion of Jews from the Austrian capital after the National Socialist takeover in 1938 either by accessing the data directly, or via a map of the city at that time which shows the concentration of Jews in any given neighborhood. For example, one can learn that until August 1, 1938, Olga Bernstein lived at Pilgerimgasse 22, at which point she was evicted from her home for being Jewish, and was subsequently deported to Minsk in what had been territory of the Soviet Union (now Belarus) on November 28, 1941, with her husband, Juda. We can also learn from the database that Bernstein’s maiden name was Fuchs, and that she was born on July 14, 1900, in the former Austrian province of Moravia. Her husband, Juda, was born in Bob rnjisk, Russia, on June 20, 1888. Also, we can see where the Bernsteins lived in the city by clicking over to the map from the database. No date of death is available from the database for either Olga or Juda, so it is unclear from the data available on this website whether either (or both) survived the war. This database, combined with its mapping capabilities, allows users to visualize not only the patterns of Jewish residence in Vienna, but also the patterns of deportation from the city over time. Students examining the map can be prompted to ask questions about the timing of the expulsions—were poor Jews expelled before wealthy Jews (based on the neighborhoods they lived in)—or was the process of clearing Jews from Vienna conducted according to some other logic? By seeing the data in geographic space, students are able to ask questions they cannot ask from the data alone.

An even more sophisticated version of this same type of historical interface is the Digital Harlem project created by Stephen Robertson of the University of Sydney. Where the NS-Crimes project only allows the user to view data from a database in geographic space, Digital Harlem lets the user take a much more active role in the creation of geographic representations of historical data. The user can specify events, people, or places, and create interactive map layers that show how these historical data map onto the geography of Harlem. Thus, a student interested in the history of prostitution in New York City could create a map layer showing arrests for prostitution and another for the locations of brothels in Harlem (fig. 8).
Right away students will see that the locations of brothels and the arrests for prostitution do not correlate very well at all. The brothels are much closer to midtown Manhattan, while the arrests for prostitution are clustered much farther uptown. This finding raises historians’ questions such as whether the police were working with the brothel owners in Harlem and so avoided arresting prostitutes close to the brothels, or conversely, if the police made the blocks around the brothels unfriendly locations for prostitutes to ply their trade, so they stayed further uptown? A

diligent student who did not have access to this map interface could puzzle out the lack of correlation between the locations of brothels and arrests for prostitution by examining the address data for each source. But with the digital interface, she can see how the historical data appear in physical space at the click of a mouse.

While these projects devoted to events in the history of Harlem and Vienna offer users a much richer experience of the historical data than they could have by simply reading sources available in various archives, these projects are still static in nature—by which I mean the user experience is delimited entirely by the website’s creators. At the other end of the spectrum is the Hypercities project created at UCLA. Using the Hypercities platform, students and/or other users can build their own interactive maps organized around a particular unit of geography—in this case, one of several cities around the world the project’s creators have made available. Once they are logged into the project and have permission to begin adding content, students can mark up the most current satellite image of their city with geotagged data—images, text, sound, or video files— which are then visible on the map via individual pins placed there. Like the Google mapping community overlays, the student-created maps in Hypercities have all the advantages and disadvantages of user-generated content. On the one hand, the items students select for inclusion on a map are reflective of their own interests, and so can be much more interesting to the students themselves. On the other hand, there is a high degree of variability in the quality of what is posted on the maps, and a number of the pins lead to “items” such as “This is where I like to jog.” Nevertheless, by handing over a fair amount of control over what is posted to the map interfaces, the creators of the Hypercities project have transferred the locus of control from the website’s creators to the website’s users—a central element of Web 2.0 interfaces. In doing so, they are turning students loose to become creators of history rather than passive consumers of history. As I have argued throughout the book, giving students this freedom to be creative is an essential element of teaching history in the digital age, but with the caveat that we must also teach them how to make the most of this freedom. Learning to make the best use of the control they are being handed—in instead of using that control to post notices about their favorite restaurants or where they jog—is something history students already need to know.

Right now, historians and history students rely on projects such as
those already described to make available limited sets of geotagged historical information. But as more and more historical sources are marked up with longitude and latitude, we can expect to see more and more and simpler and simpler interfaces for manipulating these datasets. For instance, it is already about as easy as it could possibly be to use the Yahoo Pipes platform to access geotagged data from a variety of websites. In just a two-step process it is possible to select a defined number of geotagged images from Flickr’s database that are geolocated near a particular coordinate on a map and display them as pins on a map, much like those in the Goldman project. This simple image-extraction application was created by someone (me) with virtually no knowledge of programming or computing beyond the most obvious coding needed to work with a blog or simple webpage. New applications that make this sort of mapping easier and easier appear almost monthly. The beta version of a service called “HistoryPin” offers users the opportunity to “pin their history to the world” by geotagging any historical images they own and making them available. For students of history, such services are blank slates on which they can write their own versions of the past, and it is very useful not only to let them write on that slate, but also to critique what those in the public at large have done with the interface. Are the sources others place on a map properly identified? How are they described? What can we learn (or not) from what we find on such sites? How, as historians, can we do a better job?

Text

As interesting as maps are as graphical interfaces for displaying historical data, historians still work most often with text sources, and given the amount of historical text already online and the ever-growing corpus of such text, being able to use machine methods for making sense of this massive database of historical text is no longer a luxury—it is an imperative. I have already suggested ways that historians will eventually have easy access to recombinant documents that will allow them to look for new relationships between bits or chunks of historical information. But the example of the Czech politicians relied on the historian already having an idea of what he or she was looking for; that is, evidence connected in some way to a particular meeting of political party leaders. But what happens when the historian instead confronts a database of historical data with much less well-formed questions, such as “What was the nature of the relationships
between the historical actors in this database?” or “Is there any evidence of change in family patterns over time, and if so, are those changes related at all to patterns in the economy?” or “Did Adolf Hitler’s use of anti-Semitic rhetoric vary according to the audiences for his speeches?” or “What was the impact of Spanish Jesuit missions on local economies in colonial Mexico?” These are historians’ questions—the kind we ask all the time. But finding answers to such questions, especially when those answers might require us to examine a very large amount of data, is often quite difficult and time consuming. In earlier decades, we have often narrowed the scope of our investigations to what is possible in a given amount of time given the amount of time and support we have to complete a particular project. In 1958, the one-time AHA president David H. Pinkney gave a lecture at the Newberry Library, in which he discussed why it was that American historians of France had been unable to produce magisterial studies comparable to the works of Georges Lefebvre or Albert Soboul. In his lecture, Pinkney blamed

. . . this failure to the inability of Americans, owing to geographical separation, to do the sustained work in French archives that was the foundation of the great French books. I urged my American colleagues to cease trying to meet our French friends on their own ground with monographs but instead to write on broader subjects that are of interest to Americans concerned with European history and not merely to French historians, to draw on the detailed works of others, and to study in depth in archives only neglected or debated aspects of the subject—a possible task for an American on sabbatical leave and occasional summer research trips.26

The problem Pinkney first described in 1958 has been turned on its head by digital technology. Many historians now have ready online access to too many sources on their chosen topic. Instead of worrying about how to gain access to enough sources in order to write books and articles, historians now must contend with a rapidly growing flood of sources—already so great in some cases that we cannot possibly cope with the amount of information available to us without the use of data-processing tools. Of course, this corpus of historical sources is very uneven—rich countries’ libraries and archives have been digitizing their collections at a much more rapid rate than poor countries. But the velocity of mass-digitalization proj-
Analyzing

ects is growing with each passing year. Cast your mind back ten years and recollect how many online historical sources were available in 2001, and then compare that number to what is available in 2013. Then project that growth forward another ten years, factoring in improvements in scanning techniques, and try to imagine how many online historical sources will be available to students in 2021. Even conservative projections make that number so great that the need to teach students to work with text mining and analysis software seems as obvious in 2013 as Pinkney’s advice to his colleagues was in 1958.

One simple example of how text mining can help answer historians’ questions is the matter of how to puzzle out relationships between individuals in a particular database. Text-mining algorithms are very good at this important but time-consuming task. In a series of posts in his now-defunct blog *Digital History Hacks (2005–2008)*, the Canadian historian Bill Turkel describes his use of text-mining techniques to compute such relationships in a historical database. Working with a test sample of 100 entries (from the approximately 10,000) in the *Dictionary of Canadian Biography*, Turkel used software to suggest possible relationships between his test group.27 The results were both unsurprising and surprising. As Turkel points out, he could have anticipated some of the results of this clustering analysis without the aid of the program he wrote, but other relationships suggested by the software were completely puzzling and it was the puzzling results that then required his skill as a historian to analyze. As valuable as it is to confirm what we would have already expected to learn, discovering new information in online sources that would not have been easily accessible through other means points to a significant benefit of text mining.28

In Turkel’s example, his software suggested that there is a relationship between 6 individuals in his test database of 100, but that relationship is not at all obvious at first. The only way to find out what that relationship might be is to delve directly into the data in those six entries, and it is just possible that this research effort might turn up something wholly unexpected. Given that Turkel’s program calculated all possible relationships between these 100 individuals in just a few seconds and could have done the same for all 10,000 entries in the *DCB* in around twenty-four hours, one can imagine how quickly historians will soon be able to sort through massive corpuses of text in a short amount of time.

Similarly, one could take a large body of text that is not in database form as in the prior example—for example, a novel like *Les Misérables*—
and with text-mining software determine relationships between the characters, such as how often they interact with one another. Many text-mining products allow the user to see such relationships in a graphical way that may then suggest degrees of interaction not readily apparent on the first reading of the text. Or, in the case of very large bodies of text, information presented in graphical form may make it possible for the reader to focus his or her reading of the text on only certain characters whose relationships seem to be particularly significant. Similarly, it is possible to use these same techniques to examine ideas and their relationship to one another in a corpus of text. For instance, one might take all of Adolf Hitler’s speeches during a given electoral campaign and then compare the relationships that might exist between key terms in his rhetoric such as “Jew,” “Bolshevik,” “race,” “economy,” and so on. If these speeches were then sorted by type of audience (party gathering, speech to a group of business leaders), regionally, or on the basis of size of metropolitan area, one might then be able to see whether the focus of his campaign rhetoric shifted according to audience or geographic location. Once the user can see such possible relationships, then it is possible to engage in a much more focused reading of the speeches themselves. Software still cannot analyze text in all the ways a historian would, but it can suggest interesting starting points for that analysis, and with each passing year the text mining and analysis algorithms get better and better.

What does this mean for our students and for the teaching of history in the second decade of the twenty-first century? Already, a number of simple tools exist that can be used to introduce students to the possibilities inherent in text mining. While it is not a good idea to rely solely on off-the-shelf word-cloud tools like Many Eyes or Wordle, these tools are easy to learn and can provide a useful introduction to the issues text mining raises for historians. While creating a simple word cloud from a body of text is an easy way to introduce students to the idea of text mining, these visualizations are also useful for demonstrating to them how relying on simple analysis like this can lead to erroneous conclusions like War and Peace being all about Russia. However, using this simple tool, students can be introduced to the idea of text mining by uploading a paper they have written and then playing around with the various text-visualization tools. They will see, for instance, how often they use particular words (sometimes comically so). Once introduced to text-mining techniques and the issues they raise for historical analysis, students can then be taught to use much
more sophisticated text- and data-mining engines and the visualization software that allows scholars to work with these data in more interesting and productive ways. For example, a slightly more sophisticated tool than the word-cloud packages is Google’s NGram viewer, which lets students track and compare the use of various words or phrases over time in the immense database of Google Books. As a simple example, students can track the use of “war” and “peace” in those millions of books and note that, at least in those books currently scanned by Google, “war” overtook “peace” in 1743. This finding, of course, does not mean that war was more popular than peace beginning in 1743, but rather, can point students toward productive questions about why war would be more commonly used than peace, and why by the twentieth century the difference in frequency between the two words would become so pronounced.34

Among the issues students need to be aware of is that using text mining on subtle forms of speech like political rhetoric can be a tricky proposition. Text mining works best when the text being examined by the software follows a particular set of well-defined rules. So, for instance, Dan Cohen created a simple text-mining tool he called “Syllabus Finder” in 2003 to search the Internet for course syllabi.35 Course syllabi generally follow a basic set of rules, regardless of discipline, which include text such as the professor’s name, the title of the course, the meeting pattern of the course, a course number, and things with names like “office hours,” “required readings,” etc. Using these text identifiers, Cohen was able to mine the Internet for syllabi for a number of years until Google discontinued access to the API the Syllabus Finder required.36

A political speech, however, may or may not follow a well-defined or easily discernible set of rules that makes it amenable to text mining. In the American context, for instance, oppositional terms such as “pro-choice/pro-life,” or “gun rights/gun control” may indicate the ideological position of a particular speaker, but politicians can also be much more subtle in their speech. For instance, this passage from a speech given in the U.S. Senate in 2003 by Senator Patty Murray expresses her opposition to a bill known as the “Partial-Birth Abortion Ban Act of 2003.”

Since we began debating how to criminalize women’s health choices yesterday, the Dow Jones has dropped 170 points; we are 1 day closer to a war in Iraq; we have done nothing to stimulate the economy or create any new jobs or provide any more health coverage.37
Anyone familiar with the parameters of the American debate over abortion rights will be able to tell that the phrase “debating how to criminalize women’s health choices” is a clear statement of opposition to limitations on abortion rights, but a text-mining algorithm looking for the pro-choice/pro-life pairing might well miss this particular nuance. It is certainly possible to tweak algorithms so that they produce much more sophisticated analyses of complex texts such as speeches in the U.S. Senate, but as historians come to rely more and more on such algorithms to search massive text corpuses, we will first have to learn how to do this tweaking on our own. Once we know how to do it, we will then have to figure out the best ways to teach students to do the same thing. From the simple example of Senator Murray’s speech from 2003, one can see that even with the best algorithms, historians will still need to read a certain number of primary sources in detail to make sure we have taken an inclusive view of the text identifiers the algorithm should be searching for. As mentioned earlier, anyone familiar with the parameters of the American debate on abortion rights can tell which side Senator Murray was on in 2003. But what would the text identifiers be in letters written between various representatives of the Spanish crown in South and Central America around 1800, or Chinese provincial governors around 1700? As with American political speech in the twenty-first century, the historian would need to know the letter writing conventions and the key vocabulary of Spanish or Chinese officials in order to properly instruct the algorithm as it scans all those texts. Our historian (or history student) must teach the algorithm how to search through a database of these letters, and to do that he or she must first understand that parameters of the political debate in the Spanish and Chinese empires at the time, and know how those parameters were expressed in language. Only then can text mining proceed successfully.

The ability to teach an algorithm how to search across thousands or tens of thousands of official documents more than two centuries old is, fortunately, a skill historians already possess and teach our students. We know how to make sense of these language conventions and for decades we have been teaching students how to read the same documents we read. By the end of the current decade it is a safe assumption that sophisticated data- and text-mining tools will be much more user friendly, and so therefore accessible to novice learners. If this assumption is correct, now is the time we need to develop, test, and refine teaching strategies that will incorporate these tools as they emerge. Otherwise students will either try to use
these tools on their own with limited or mixed results, or, more likely, will not use them at all, and the degree they receive will be ever more outdated.

Image Mining

The other large category of historical sources that historians rely on is images. Sorting through the seemingly limitless databases of historical images is currently a very inefficient process. The user must either use a search engine such as the Google or Yahoo image search, which returns images in an order that is not particularly useful, or must already know which database to search through to find what he or she is looking for (e.g., American Memory). In either case, the student conducting the search is dependent upon the metadata added to images for either type of search to work at all. Because the object of data mining is to turn up new information not readily available in other ways and to provide analysis of that information, this sort of image browsing does not qualify as “image mining.”

Mining visual sources for usable information is much trickier than the mining of text for several reasons. The most important of these is that while text follows the sorts of rules discussed earlier (grammar, structure of the text, etc.), images follow very few rules that can be used in historical data mining. Among the few objective bits of information common to all images are size of the image and the makeup of the pixels in that digital image; that is, how many blue, how many red, and what the density of those pixels is in any particular quadrant of the image. These sorts of basic data provide some information that is usable for humanists, and certainly even this limited amount of data will lead to the creation of new knowledge about the content of the images.39 For now, though, we lack clear intersections between the underlying data—size, pixels—that can be extracted from the image and the meanings that can be made from interpretation of the content of the image, sometimes known as the “semantic gap.”40 This gap in meaning making—one that those working in the field of text mining are beginning to bridge already—is really no more than an engineering problem that will be overcome soon enough. This particular engineering problem is more difficult to deal with because we do not yet even have a reliable way to locate images that are related to one another across multiple databases absent metadata providing those links. However, software designers are beginning to make progress when it comes to this latter task. It is already possible to train a search algorithm to ferret out
images of a particular object—a motorcycle, for instance—by determining which sectors of the image of a motorcycle might be indicative of images of any motorcycle. Once several such sectors have been identified, then the algorithm can assume that any image it scans that contains a sufficient number of matches for those sectors must be (or at least is likely to be) a motorcycle. Already this technology is being used to combat online child pornography by identifying images that might include not only children, but sexual content.

This very rudimentary process of identifying objects such as motorcycles is but the first step toward a much more robust capability to search databases of historical images. Imagine for a moment what it will be like when a student working on a paper on the diffusion of steam technology in the nineteenth century is able to search across a cluster of large databases of historical images for possible images of a particular model of steam engine. If the software is sufficiently robust, it might also be possible to identify different models of the same engine based upon unique characteristics of the engine itself, if such details are present in an image. Depending on the metadata available for the various images returned in such a search (date, location, image creator, etc.) it may well be possible to do such things as map out the locations of these steam engines and the dates they were photographed. Seeing the diffusion of this particular technology over time and space may suggest new questions, new answers, or simply new avenues for investigation to students. Or, instead of an industrial product like a steam engine, what if students were working on images of a particular public figure (artist, politician, social reformer) and could use the software to ferret out all images of that figure? What might be learned from such information? What new questions might be generated? Or what if a student was interested in the use of a particular image in books, magazines, and digital media? Take for instance, an iconic photograph like Dorothea Lange’s Migrant Mother, which appears in hundreds, if not thousands, of books and articles, and on countless websites. Because Lange’s 1936 image of Florence Owens Thompson and her children is a singular item, searching algorithms can be trained to locate this item with much greater ease, and can return such additional data as the title of the book where the image was located, the page number, author, date of publication, and so on. We are still a way off from mining images in these ways, but given what is already possible with existing algorithms, these same scenarios might be possible in as little as five years. Given that students may very well be able to engage
in this kind of image mining soon, it is incumbent upon us as educators to begin working on ways to train them to do this sort of sophisticated work. How will our students survive and prosper as historians in a world with millions of books, and billions of other sources available online at the click of a mouse? They will do so only if historians begin to take seriously the need to train students to work with not only the vast quantities of historical information now available to them, but also with the increasingly sophisticated software tools under development for working with those resources. To do that, of course, we have to learn to use these tools ourselves so that we can develop useful models for students: teaching and learning exercises that help them make sense of the huge online library of historical resources. Finally, we need to begin thinking carefully as a community of scholars about the kinds of historical questions one can reasonably ask of these super-massive databases. Once we have a better handle on what those questions are and how we might go about answering them, then we can engage students in a lively discussion about both the questions and the possible answers. Because students often have technology skills that are substantially greater than our own, inviting them to be part of this discussion will almost certainly be well worth the effort.