

The Hermeneutics of Data and Historical Writing

by Frederick W. Gibbs and Trevor J. Owens

Ongoing digitization of primary sources and the proliferation of born-digital documents are making it easier for historians to engage with vast amounts of research material. As a result, historical scholarship increasingly depends on our interactions with data, from battling the hidden algorithms of Google Book Search to text mining a hand-curated set of full-text documents. Even though methods for exploring and interacting with data have begun to permeate historical research, historians' writing has largely remained mired in traditional forms and conventions. This chapter discusses some new ways in which historians might rethink the nature of historical writing as both a product and a process of understanding.

We argue that the new methods used to explore and interpret historical data demand a new level of methodological transparency in history writing. Examples include discussions of data queries, workflows with particular tools, and the production and interpretation of data visualizations. At a minimum, historians need to embrace new priorities for research publications that explicate their process of interfacing with, exploring, and then making sense of historical sources in a fundamentally digital form—that is, the hermeneutics of data.¹ This may mean de-emphasizing narrative in favor of illustrating the rich complexities between an argument and the data that supports it. It may mean calling attention to productive failure--when a certain methodology or technique proved ineffective or had to be abandoned. These are precisely the kinds of lessons historians need to learn as they grapple with new approaches to making sense of the historical record.

In this essay we consider data as computer-processable information. This includes measurements of nearly every kind, such as census records, as well as all types of textual publications that have been rendered as plain text. We must also point out that, while data certainly can be employed as evidence for a historical argument, data are not necessarily evidence in themselves. Nor do we consider data necessarily to be a direct representation of the historical record, as they are also produced by tools used to investigate or access large datasets. Given the myriad forms that data can take, making sense of data and using

¹ The authors prefer to use "data" as a singular mass noun, referring not to multiple datum, but to the historical record that can be represented as digital text. Such usage parallels our use of "text" as comprising all relevant texts.

them as evidence has become a rather different skill for historians than it has been. For that reason, we argue that the creation of, interaction with, and interpretation of data must become more integral to historical writing.

We call upon historians to publicly experiment with ways of presenting their methodologies, procedures, and experiences with historical data as they engage in a cyclical process of contextualization and interpretation. This essay hopes to encourage more dialog about why historical writing must foreground methodological transparency and free itself from the epistemological jitters that make many historians wary of moving away from close readings or embracing the notion of the historical record as data.

Data in History

Use of data in the humanities has recently attracted considerable attention, and no project more so than Culturomics, a quantitative study of culture using Google Books.² Of course the idea of using data for historical research is hardly new, whether in the context of quantitative history, early work from the Annales school, or work done under the rubric of humanities computing. Yet the nature of data and the way it has been used by historians in the past differs in several important respects from contemporary uses of data. This is especially true in terms of the sheer quantity of data now available that can be gathered in a short time and thus guide humanistic inquiry. The process of guiding should be a greater part of our historical writing.

Some scholars who work within the domain of the digital humanities have begun to think and write more explicitly about data and its potential for new kinds of research. For example, some Shakespeare scholars have been using statistical procedures to identify language features that signal classification in dramatic works.³ The [Stanford Literary Lab](#) has been rethinking the nature of genre through semantic analysis. Yet most projects, including these, continue to be largely confirmatory, like reinforcing the periodization of Shakespeare's plays or confirming the codified family of literary genres. To be clear, this is not a criticism of these projects and their outcomes---they are in fact a crucial step forward. As humanists continue to prove that data manipulation and machine learning can confirm existing knowledge, such techniques come closer to telling us something we don't already know. Other large scale research projects, like those funded through the [Digging Into Data Initiative](#), have begun to explore the transformative potential of data in

² Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, et al. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331.6014 (2011): 176 -182.

³ Michael Witmore and Jonathan Hope, "Shakespeare by the Numbers: On the Linguistic Texture of the Late Plays," in *Early Modern Tragicomedy*, ed. Subha Mukherji and Raphael Lyne. (Cambridge: D.S.Brewer, 2007).

humanities research as well.⁴

However, even these projects generally focus on research (or research potential) rather than on making their methodology accessible to a broader humanities audience. To some extent, legitimizing digital work does require an appeal to the traditional values (and forms) of the non-digital humanities. But how can digital historians expect others to take their new methodologies seriously when new ways of working with data (even when not with sophisticated mathematics) remain too much like an impenetrable and mysterious black box? The processes for working with the vast amounts of easily accessible and diverse large sets of data suggest a need for historians to formulate, articulate, and propagate ideas about how data should be approached in historical research.

Towards a Hermeneutics of Data

What does it mean to "use" data in historical work? To some extent, historians have always collected, analyzed, and written about data. But having access to vastly greater quantities of data, markedly different kinds of datasets, and a variety of complex tools and methodologies for exploring it means that "using" signifies a much broader range of activities than it has previously. The rapid rate of data production and technological change means that we must continue to teach each other how we are using and making sense of data.

We should be clear about what using data does *not* imply. For one, it does not refer only to historical analysis via complex statistical methods to create knowledge. Even as data become more readily available and as historians begin to acquire data manipulation skills as part of their training, rigorous mathematics is not necessarily essential for using data efficiently and effectively. In particular, work with data can be exploratory and deliberately without the mathematical rigor that social scientists must use to support their epistemological claims. Using data in this way is fundamentally different from using data for quantifying, computing and creating knowledge as per quantitative history.

Similarly, historians need not treat and interpret data only for rigorous hypothesis testing. This is another crucial difference between our approach and the approaches of the cliometricians of the 1960s and 70s.⁵ Perhaps such a potential dependence on numbers

⁴ Stanford Literary Lab, <http://litlab.stanford.edu/>; Digging into Data Challenge, Office of Digital Humanities, National Endowment for the Humanities, <http://www.diggingintodata.org/>.

⁵ For a more detailed history of cliometrics and its impact on the digital humanities see William G Thomas III, "Computing and the Historical Imagination," in *A Companion to Digital Humanities*, ed. Susan Schreibman, Raymond George Siemens, and John M. Unsworth. (Oxford: Blackwell, 2004).

became even more unpalatable to non-numerical historians after an embrace of the cultural turn, the importance of subjectivity, and a general epistemological stance against the kind of positivism that underpins much of the hypothesis testing baked into the design of statistical procedures and analytical software.

But data does not always have to be used as evidence. It can also help with discovering and framing research questions. Especially as increasing amounts of historical data is provided via, or can be viewed with, tools like Google's Ngram Viewer (to take a simple example), playing with data—in all its formats and forms—is more important than ever. This view of iterative interaction with data as a part of the hermeneutic process---especially when explored in graphical form---resonates with some recent theoretical work that describes knowledge from visualizations as not simply “transferred, revealed, or perceived, but...created through a dynamic process.”⁶ Data in a variety of forms can provoke new questions and explorations, just as visualizations themselves have been recently described as “generative and iterative, capable of producing new knowledge through the aesthetic provocation.”⁷

As the investment of time and energy to acquire data decreases, rapidly working with data can now be a part of historians' early development and exploration of a research question. So too can it quickly illustrate potentially interesting but ultimately dead-ends of scholarly research---“negative results,” perhaps, that should not be discarded as they likely would be for a typical scholarly book or journal article. It bears repeating that using large amounts of data for research should not be considered opposed to more traditional use of historical sources. As historical data become more ubiquitous, humanists will find it useful to pivot between distant and close readings. More often than not, distant reading will involve (if not require) creative and reusable techniques to re-imagine and re-present the past---at least more so than traditional humanist texts do. For this very reason, it becomes insufficient to simply write about research as if it's independent of its methodology.

Furthermore, rich datasets (like the National Archives' [Access to Archival Databases](#)) and interfaces to data (like [Google Fusion Tables](#)) are making it easier than ever for historians to combine different kinds of datasets—and thus provide an exciting new way to triangulate historical knowledge.⁸ Steven Ramsay has suggested that there is a new kind of role for searching to play in the hermeneutic process of understanding,

⁶ Martin Jessop, “Digital Visualization as a Scholarly Activity,” *Literary and Linguistic Computing*, 23.3 (2008): 281-293, 282.

⁷ Johanna Drucker, “Graphesis: Visual Knowledge Production and Representation,” *Poetess Archive Journal*, 2.1 (2010), <http://paj.muohio.edu/paj/index.php/paj/article/view/4>.

⁸ Access to Archival Databases, The National Archives, <http://aad.archives.gov/aad/>; Google Fusion Tables, <http://www.google.com/fusiontables/public/tour/index.html>.

especially in the [value of ‘screwing around’](#) and embracing the serendipitous discovery that our recent abundance of data makes possible.⁹ This could result, for example, in noticing within the context of London’s central criminal court, [the Old Bailey](#), that [trials about poisoning tend to refer to coffee](#) more than to other beverages, and very rarely to food.¹⁰ Thus, our methodologies might not be as deliberate or as linear as they have been in the past. And this means we need more explicit and careful (if not playful) ways of writing about them.

Methodology in Writing

Despite some recent methodological experimentation with data, historians have not been nearly as innovative in terms of writing about how they use it. Even as scholars (at least in certain fields) have embraced communication with new media, historical writing has been largely confined by linear narratives, usually in the form of journal articles and monographs. The insistence on creating a narrative in static form, even if online, is particularly troubling because it obscures the methods for discovery that underlie the hermeneutic research process.

Historical work has needed to tell a good story, but methodology has not made for a very good story or the kind of historical writing that is likely to be published in traditional venues. Although relatively simple text searches or charts that aid in our historical analysis are perhaps not worth including in a book, our searches and work with data have grown increasingly complex, as has the data available to us. While these can present new perspectives on the past, they can only do so to the extent that other historians feel comfortable with the methodologies that are used. This means using appropriate platforms to explain our methods. Does it make sense to explain new research methods that are wholly dependent on large datasets and their manipulation and visualization in a static book that distances the reader from the tools and techniques being described? Of course the realities of the profession restrict publishing freedoms (no one has gotten tenure for a really good website version of their dissertation), but our work need not be restrained by a false dichotomy between new media and old media. We suggest that exploratory methodological work can exist online in a perfectly complementary way to more traditional publication venues---and that the symbiotic pair will make both elements the

⁹ Stephen Ramsay, “The Hermeneutics of Screwing Around; or What You Do with a Million Books,” (Conference paper presented at Playing With Technology in History, 2010), <http://www.playingwithhistory.com/wp-content/uploads/2010/04/hermeneutics.pdf>.

¹⁰ The Proceedings of the Old Bailey, 1674-1913, <http://www.oldbaileyonline.org>; Frederick Gibbs, “Beware the Coffee,” *With Criminal Intent* (March 29, 2011), <http://criminalintent.org/2011/03/beware-the-coffee/>.

better for it.

Regardless of form, we need history writing that explicates the research process as much as the research conclusions. We need history writing that interfaces with, explains, and makes accessible the data that historians use. We need history writing that will foreground the new historical methods to manipulate text/data coming online, including data queries and manipulation, and the production and interpretation of visualizations. As John Unsworth suggested long ago with respect to hypertext projects, history writing should explicate failure wherever possible.¹¹ As Tim Sherratt and Bethany Nowviskie suggested in their comments on an earlier version of this chapter, one inspiring model for a new kind of publication is the artist's sketchbook that maps out ideas, explorations, false starts, and promising leads.¹²

There is no question that humanists can be---and in fact are trained to be---skeptical of data manipulation. This is perhaps the preeminent reason why methodology needs to be, at least for now, clearly explained. With new digital tools, we are still groping to understand how to identify the best methods for very messy circumstances of historical data. However, the reasons why many historians remain skeptical about data are not all that different from the reasons they can be skeptical about text. Historians have long reflected on the theoretical advantages and practical limitations of various methodologies and approaches to textual research. Critical theorists and historians alike have commented on the slippery notion of a text; some excellent theoretical work on cybertext and hypertext have muddied the waters further. The last few years have complicated such a notion even more as many traditional texts have come to be seen as data that can be quickly searched, manipulated, viewed from a variety of perspectives, and combined with other data to create entirely new research corpora. Just as the problematic notion of a text has not undermined the hermeneutic process, nor should the notion of data. It is clear that a new relationship between text and data has begun to unfold.¹³ This relationship must inform our approach to writing as well as research.

One way of reducing hostility to data and its manipulation is to lay bare whatever manipulations have led to some historical insight. Methodological tutorials, for example, would not only help legitimate the knowledge claims that employ them, but make the methodology more accessible to anyone who might recognize that the same or slightly modified approach could be of value in their own work. Beyond explicit tutorials, there

¹¹ John Unsworth, "Documenting the Reinvention of Text: The Importance of Failure," *The Journal of Electronic Publishing* 3.2 (1997), <http://dx.doi.org/10.3998/3336451.0003.201>.

¹² Comments by [Tim Sherratt](#) and [Bethany Nowviskie](#) on "The Hermeneutics of Data and Historical Writing," *Writing History in the Digital Age*, web-book edition, Fall 2011.

¹³ Julia Flanders, "Data and Wisdom: Electronic Editing and the Quantification of Knowledge," *Literary and Linguistic Computing* 24.1 (2009): 53-62.

are several key advantages in foregrounding our work with data: 1) It allows others to verify historical claims; 2) It is instructive as part of teaching and exposing historical research practices; 3) It allows us to keep pace with changing tools and ways of using them. Besides, openness has long been part of the ethos of the humanities, and humanists continually argue that we should embrace more public modes of writing and thinking as a way to challenge the kind of work that scholars do. For example, Dave Perry's blog post, "[Be Online or Be Irrelevant](#)," suggests that academic blogging can encourage "a digital humanism which takes down those walls and claims a new space for scholarship and public intellectualism."¹⁴ This cannot happen unless our methodologies with data remain transparent.

Case Study: Becoming Users and Communities of Data

Our theoretical and prescriptive remarks thus far will benefit from a concrete example: in this case, one that explores the history of the user. The notion of the user has become ubiquitous. We live in an era of usernames, user experiences, and user-centered design; we tacitly sign end-user license agreements when we install software; we read user guides to figure out how to get our software to do what we want. But our omnipresent conception of ourselves as users obscures the history of the term.

Of course, it now takes only seconds to follow this line of inquiry (the history of a term), and see the relationship between the presence of that term and any other similar terms in [Google's Ngram Viewer](#), which allows anyone to chart the frequency of words or phrases across a subset of the digitized Google Books corpus.¹⁵

¹⁴ Dave Perry, "Be Online or Be Irrelevant: Thoughts on Emerging Media and Higher Education," *AcademHack*, January 11, 2010, <http://academhack.outsidethetext.com/home/2010/be-online-or-be-irrelevant/>.

¹⁵ Google Books Ngram Viewer, <http://books.google.com/ngrams>.

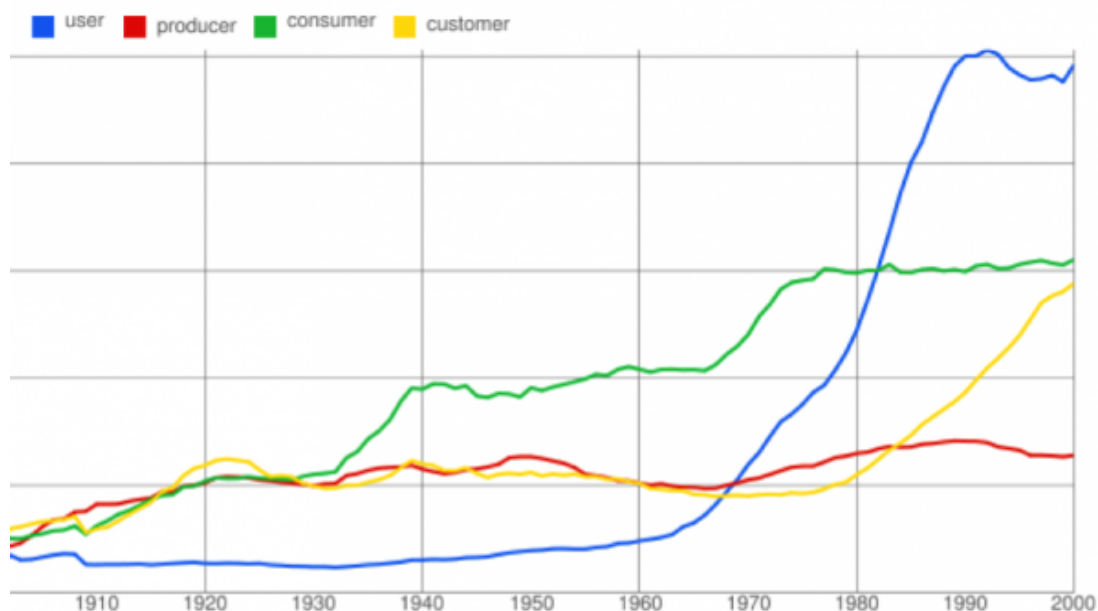


Figure 1: Frequency of selected words (user, producer, consumer, customer), 1900-2000, from Google Books Ngram Viewer.

Needless to say, this chart is not historical evidence of sufficient (if any) rigor to support historical knowledge claims about what is or isn't a user. For one, Google's data is proprietary and exactly what comprises it is unclear. Perhaps more importantly, this graph does not indicate anything interesting about *why* the term "user" spiked as it did—the real question that historians want to answer. But these are not reasons to discard the tool or to avoid writing about it. Historians might well start framing research questions this way, with quick uses of the Ngram viewer or other tools. Conventionally, this work would remain invisible, and only "real" data would appear in published work to support an argument of influence or causation. But foregrounding such preliminary work (like Ngram charts) will help readers to understand the genesis of the question, flag possible framework errors, identify category mistakes, and perhaps inspire them to think about how such techniques might benefit their own work.

To investigate the user in more detail, one can use other online corpora to generate a series of radically different interpretive views. For example, searching in the [Time Magazine Corpus](http://corpus.byu.edu/time/) allows one to see all of the collocates (words that appear within a specified number of words from the search term) and display counts by decade.¹⁶

¹⁶ Time Magazine Corpus, Brigham Young University, <http://corpus.byu.edu/time/>.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Category	Collocate	Frequency	20s	30s	40s	50s	60s	70s	80s	90s	00s
2	Drugs	DRUG	32		1			3	6	11	7	4
3	Drugs	COCAINE	6						2	1	3	
4	Drugs	MARIJUANA	6					1	5			
5	Drugs	HEROIN	5					2	2	1		
6	Drugs	ALCOHOL	4	1				1			1	1
7	Drugs	LSD	3					2	1			
8	Drugs	CRACK	3							3		
9	Technology	FRIENDLY	13							7	2	4
10	Technology	INTERFACE	10							1	2	7
11	Technology	COMPUTER	9							3	5	1
12	Technology	WINDOWS	7								5	2
13	Technology	TELEPHONE	6				3	2	1			
14	Technology	PHONE	5						2	3		
15	Technology	INTERNET	4								1	3
16	Technology	NAMES	4								3	1
17	Technology	IDS	3								2	1
18	Technology	MACINTOSH	3								1	2
19	Technology	ID	3								2	1
20	Technology	AOL	3								2	1
21	Technology	PC	3							1	2	
22	Technology	TOOL	3						1		2	
23	Technology	SWITCH	3						2			1
24	Technology	TOOLS	3		1	1				1		
25	Technology	MAKER	3			2	1					
26	Technology	COMPUTERS	3							3		
27	?????	STEEL	13			3	9	1				
28		FEES	8							4	3	1
29		FEE	4				1			3		
30		TAXES	4				3	1				
31		TIN	3	1	2							
32		PRODUCER	3		1	1	1					
33		LANGUAGE	3					1		1		1

Figure 2: What are users using? Frequency of collocates of “user” by decade, 1920s-2000s, from Time Magazine Corpus.

Column B in Figure 2 lists most words that appear within a four-word window of “user.” So it's easy to see, for example, that “drug” appears within 4 words of “user” 32 times. To better make sense of these results, the collocates were coded into two categories (column A): those that have to do with drugs and those that have to do with technology. There are a few at the bottom that remain uncategorized, but which most likely would be considered technology uses of the term. To draw attention to the patterns in the data, cells in the sheet with two hits have been coded dark green; those with more than two hits, light green.

On the whole, this chart about “users” lends itself to some quick observations. For example, as far as these keywords in the Time Magazine Corpus suggest, the growth around the term “user” happened for drugs a bit earlier than for technology, although the latter context came to be the predominant one. We can also see that one of the first technology terms to appear is “telephone,” which perhaps suggests that the rise of the “user” may have less to do with the rise of computing (our typical conception of it now) than the rise of networks.

But going beyond the data—making sense of it—can be facilitated by additional expertise in ways that our usually much more naturally circumscribed historical data has generally not required. [Owens blogged about this research](#) while it was in progress, describing what he was interested in, how he got his data, how he was working with it, along with a link for others to explore and download the data.¹⁷ Over the next week the post was viewed over two hundred times; twenty-two researchers and librarians tweeted about the post. Most importantly, Owens received several substantive comments from scholars and researchers. These ranged from encouraging the exploration of technical guides, learning from scholarship on the notion of the reader in the context of the history of the book, and suggestions for different prepositions that could further elucidate semantic relationships about “users.” This discussion resulted from Owens having foregrounded online his initial forays into data where it was easy to give different views of his data. Sharing preliminary representations of data, providing some preliminary interpretations of them, and inviting others to consider how best to make sense of the data at hand, quickly sparked a substantive scholarly conversation. This is not to say we should expect everyone to help with our own research. Rather, that because we have so much raw data that ranges widely over typical disciplinary boundaries, a collaborative approach is even more essential to making sense of data. And it benefits everyone involved, as the discussants can learn about data and methodologies that might be useful in their own work.

In addition to accelerating research, foregrounding methodology and (access to) data gives rise to a constellation of questions that are becoming increasingly relevant for historians. How far, for example, can expressions of data like Google’s Ngram viewer be used in historical work? Although a chart from historical data should not be automatically admitted as historical evidence in itself, it certainly *can* be used to identify curious phenomena that are unlikely to be artifacts of the data or viewer alone. But how does one cite data without black-boxy mathematical reductions, and bring the data itself into the realm of scholarly discourse? How does one show, for example, that references to “sinful” in the nineteenth century appear predominantly in sermon and other exegetical literature in

¹⁷ Trevor Owens, “When Did We Become Users?” August 5, 2011, <http://www.trevorowens.org/2011/08/when-did-we-become-users/>.

the early part of the century, but become overshadowed by more secular references later in the century? Typically, this would be illustrated with pithy, anecdotal examples taken to be representative of the phenomenon. But does this adequately represent the research methodology? Does it allow anyone to investigate for themselves? Or learn from the methodology?

Far better would be to explain the steps used to collect and reformat the data; ideally, the data would be available for download. The plain text file that was reformatted to make the above linguistic shift in "sinful" detectable would be considerably useful for other researchers, who in turn will certainly make other observations and draw new and perhaps contradictory conclusions. Exposed data allow us to approach interesting questions from multiple and interdisciplinary points of view in the way that citations to textual sources do not. Again, this is not to argue for a whole-cloth replacement of close readings and textual analysis in historical research, but rather for complementing them with our explorations of data. As it becomes easier and easier for historians to explore and play with data it becomes essential for us to reflect on how we should incorporate this as part of our research and writing practices. Is there a better way than to simply provide the raw data and an explanation of how to witness the same phenomenon? Is this the twenty-first century footnote?

Conclusions

Overall, there has been no aversion to using data in historical research. But historians have started to use data on new scales, and to combine different kinds of data that range widely over typical disciplinary boundaries. The ease and increasing presence of data, in terms of both digitized and increasingly born digital research materials, mean that---irrelevant of historical field---the historian faces new methodological challenges. Approaching these materials in a context sensitive way requires substantial amounts of time and energy devoted to how exactly we can interpret data. Consequently, we have argued that historians should deliberately and explicitly share examples of how they are finding and manipulating data in their research with greater methodological transparency in order to promote the spirit of humanistic inquiry and interpretation.

We have also argued that working with and writing about data does not mean that historians need to shoulder the kinds of epistemological burdens that underpin many of the tools that statisticians or quantitative historians have developed. This is not to say that statistics are not a useful tool for inquiry, but rather that the mere act of working with data does not obligate the historian to rely on abstract data analysis. Historical data might require little more than simple frequency counts, simple correlations, or reformatting to make it useful to the historian looking for anomalies, trends, or unusual but meaningful coincidences.

To argue against the necessity of mathematical complexity is also to suggest that it is a mistake to treat data as self-evident or that data implicitly constitute historical argument or proof. Historians must treat data as text, which needs to be approached from multiple points of view and as openly as possible. Working with data can be playful and exploratory, and useful techniques should be shared as readily as research discoveries. While typical history scholarship has largely kept methodology and data manipulation in the background, new approaches to writing can complement more traditional methods and venues to avoid some of their well documented limitations, especially as it enables sharing data in a variety of forms.

Gathering data, manipulating it, representing it, and of course writing about it, should be required of all historians in training---not just those in digital history or new media courses---to best use the new kinds of historical data that have opened up new avenues of inquiry for virtually every historical specialty. Of course not all research projects will require facility with data. But just as historians learn to find, collect, organize, and make sense of the traditional sources, they also need to learn to acquire, manipulate, analyze, and represent data. Access to historical sources makes the historical record in the twenty-first century looks rather different than it has ever before. Writing about history needs to evolve as well.

About the authors: [Fred Gibbs](#) is an assistant professor of history at George Mason University and the director of digital scholarship at the Roy Rosenzweig Center for History and New Media. [Trevor Owens](#) is a digital archivist at the Library of Congress. He also teaches digital history at American University.