EDITORIAL THEORY AND LITERARY CRITICISM
George Bornstein, Series Editor

Palimpsest: Editorial Theory in the Humanities,
edited by George Bornstein and Ralph G. Williams

Contemporary German Editorial Theory, edited by Hans Walter Gabler,
George Bornstein, and Gillian Borland Pierce

The Hall of Mirrors: Drafts & Fragments and the End of Ezra Pound's Cantos,
by Peter Stoicheff

Emily Dickinson's Open Folios: Scenes of Reading, Surfaces of Writing,
by Marta L. Werner

Scholarly Editing in the Computer Age: Theory and Practice, Third Edition,
by Peter L. Shillingsburg

The Literary Text in the Digital Age, edited by Richard J. Finneran

The Margins of the Text, edited by D. C. Greetham

A Poem Containing History: Textual Studies in The Cantos,
edited by Lawrence S. Rainey

Much Labouring: The Texts and Authors of Yeats's First Modernist Books,
by David Holdeman

Resisting Texts: Authority and Submission in Constructions of Meaning,
by Peter L. Shillingsburg

The Iconic Page in Manuscript, Print, and Digital Culture,
edited by George Bornstein and Theresa Tinkle

Collaborative Meaning in Medieval Scribal Culture: The Otho Laȝamon,
by Elizabeth J. Bryan

Oscar Wilde's Decorated Books, by Nicholas Frankel

Managing Readers: Printed Marginalia in English Renaissance Books,
by William W. E. Slights

The Fluid Text: A Theory of Revision and Editing for Book and Screen, by John Bryant

Textual Awareness: A Genetic Study of Late Manuscripts by Joyce, Proust, and Mann,
by Dirk Van Hulle

The American Literature Scholar in the Digital Age, edited by Amy E. Earhart and
Andrew Jewell

DIGITALCULTUREBOOKS is an imprint of the University of Michigan Press and the Scholarly Publishing Office of the University of Michigan Library dedicated to publishing innovative and accessible work exploring new media and their impact on society, culture, and scholarly communication.

The American Literature Scholar in the Digital Age

EDITED BY

Amy E. Earhart and Andrew Jewell

THE UNIVERSITY OF MICHIGAN PRESS AND
THE UNIVERSITY OF MICHIGAN LIBRARY
ANN ARBOR

used only eight times and describe national borders or other literal border markers, like creeks in a pasture or conch shells in a garden. The other variations on *border* in the list do not appear at all in the prairie novels. Cather's metaphorical use of a variation of *border* is unique to "The Sculptor's Funeral."

[16] Willa Cather to Roscoe Cather, [November 28, 1918], Roscoe and Meta Cather Collection, Archives and Special Collections, University of Nebraska-Lincoln Libraries.

[17] Currently, however, I am aware of only a handful of scholars who use text analysis in their literary criticism. See, e.g., Tanya Clement, "'A thing not beginning and not ending': Using Digital Tools to Distant-Read Gertrude Stein's *The Making of Americans*," *Literary and Linguistic Computing* 23, no. 3 (2008): 362; the work of David L. Hoover, such as "The Future of Text Analysis" (keynote address, Canadian Symposium on Text Analysis, Saskatoon, Saskatchewan, Canada, 17 October 2008); or the dissertation work of Sarah Steger at the University of Georgia.

Visualizing the Archive

EDWARD WHITLEY

During the summer of 2008, *Time* magazine ran an article about the negative effects of digital media on human society. I have read a lot of articles like this over the years. Some of them have a profound effect on me, forcing me to rethink my reliance on word processors and contemplate a return to longhand composition. I am able to scoff at other such articles as the ill-founded fears of Luddites and technophobes. This particular article, however, left me feeling neither frightened nor smug but, instead, made me stop and reflect on what it is that the digital literary archives I have spent much of my professional life concerned with actually *do* that makes them better (or even different) than their print counterparts. The most arresting moment in this article was the suggestion that the centuries-old medium of the printed newspaper offers something that the digital revolution has struggled (if not outright failed) to provide, something that the article describes as an intellectual process of "serendipitous discovery and wide-angle perspective."[1]

I have read enough newspapers to think I know what this means: when you fold open a page of newsprint on your dining-room table, you have before you a series of hyperlinked texts in a visual arena much larger than even the biggest computer monitor, and as your eye is drawn from one article to the next, you find your perspective broadened through a series of unexpected discoveries. For as long as I have been working with electronic archives of American literature, though, I have thought that the real advantage that the digital medium has over print is that a rich archive of electronic texts can offer a "wide-angle perspective" on a large body of material,

material that is then searchable in ways that allow for the "serendipitous discovery" of new knowledge. Maybe it is naive of me to want the digital medium to be exponentially better, faster, and more sophisticated than its print predecessor, but if *Time* magazine is right that printed newspapers have already been providing "serendipitous discovery and wide-angle perspective" for centuries, then those of us who work with digital archives are not doing as much as we think we are to exploit the unique properties of the medium.

In this essay, I consider some of the opportunities that scholars working with digital archives have at their disposal for using the electronic medium to study literature in ways that would be difficult (if not impossible) to duplicate in print. Specifically, I look at digital text visualization tools, such as tools that display word patterns in graphical format and tools that rearrange the words of a text into playful and thought-provoking images. These visualization technologies not only have the potential to transform how we currently use digital literary archives, but they also challenge us to read texts differently than we otherwise would. At present, digital literary archives are rich, if somewhat static, repositories of information that give scholars and students more or less two methods for working with the documents they house: browse mode and search mode.[2] In browse mode, digital archives allow for a wide-angle perspective on their material by trusting to the wanderings of a curious mouse clicker. In search mode, the hope is that a search engine will serendipitously discover information that a browsing scholar or student might otherwise miss. Browse mode shows the patterns of the forest, while search mode pinpoints specific trees. The database structure that underlies many digital literary archives (either literally or figuratively) is designed to produce precisely this effect. As Stephen Ramsay writes, "To build a database one must be willing to move from the forest to the trees and back again. . . . [T]o use a database is to reap the benefits of enhanced vision from which the system affords."[3]

But the "enhanced vision" that Ramsay rightly attributes to the structure of many digital archives is still, at present, limited. By taking greater advantage of the visualization tools that scholars and professionals in the fields of computer science, graphic design, and information architecture have developed in recent years, those of us who work with digital archives will have the opportunity not only to enhance our vision but also to rethink some of our basic assumptions about how to read.[4] Most text visualization tools carry with them a number of methodological and theoretical implications about reading that run counter to what scholars and teachers of literature traditionally think of as the "proper way" to read a text. Rather than ask us to perform a close reading of a text (as we might ordinarily do), text visualization tools propose *distant reading* and *spatial reading* as complementary practices to the traditional method of close reading, practices that promise, among other things, wide-angle perspective on the large corpora of texts housed in digital archives and serendipitous discovery of the knowledge these archives contain.[5]

Distant Reading

Most of us are familiar with visual representations of numerical data. Pie charts, bar graphs, and scatter plots appear frequently in newspapers and textbooks and on the evening news. Such visualizations help us to perceive patterns in data that we might otherwise miss and to hear the stories that numbers alone might otherwise struggle to tell. Because numbers are quantitative, turning numerical data into a visual image is a relatively straightforward task. But words, which are more qualitative than quantitative, are another matter (and the words of literary texts, which I will get to in a moment, are yet another matter entirely). During election season, we are often surfeited with pie charts and bar graphs as news outlets present us with information graphics that attempt to reduce voter opinion to a single image that is then made to serve as a representative snapshot of the nation as a whole. But even though these ubiquitous campaign infographics are based on numerical data, those data were originally collected through verbal conversations between pollsters and voters, conversations that were rich in nuance and detail. As anyone who has ever fielded a phone call from a pollster knows, conversations that begin with open-ended questions like "Are you better off than you were four years ago?" invariably end with questions that attempt to turn your words into a quantifiable number, such as "On a scale of one to ten, are you better off than you were four years ago?" It is the job of pollsters to turn detailed conversations into raw numbers, and then those numbers—not the original words—determine the shape of the information graphic. For literature scholars, however, words *are* data, not static noise that needs to be winnowed away to get at the quantifiable information that can then be plotted on a visual graph.

Given that the entire profession of information visualization has grown up around quantifiable data, it comes as no surprise that literature scholars have been reluctant to turn to graphical representation as a methodology for interpreting literary texts. Literature scholars tend to value close reading—the subtlety of word choice, the nuance of phrasing—over the broad brush-strokes of information visualization. In many ways, the two fields seem to be at a methodological impasse: the virtue of information visualization is that it can make complex data sets more accessible than they might otherwise be, whereas literary close readings often reveal that apparently straightforward texts are more complex than they might otherwise seem. Information visualization seems better suited to analyzing the ups and downs of marketing trends or the changing patterns of crime in a big city than to interpreting the language of literary texts. Nevertheless, scholars working in the digital humanities have found ways to use the tools of information visualization to supplement traditional close readings of literary texts. Instead of parsing out the nuance of individual words and phrases, these scholars have used digital technology to search for patterns and to trace broad outlines, either in a single text or across a body of related texts.

A number of these scholars have cited Franco Moretti's concept of "distant reading" in an effort to differentiate between the traditional practice of close reading and the new ways that digital technologies are allowing literature scholars to read texts. Moretti has argued that close reading of individual texts is not the best way to keep track of the thousands of texts that make up literary history. Rather, he counsels scholars to step back and look at the broad patterns that emerge when you consider a wide swath of texts. He writes that "instead of concrete, individual works" serving as the building blocks of literary history, large-scale patterns of publication and reception provide "a sharper sense of [the] overall interconnection" of texts.[6] Moretti's 2005 book *Graphs, Maps, Trees: Abstract Models for a Literary History* is filled with information graphics that detail, for example, the rise of the British novel from 1700 to 1840 and the number of European book imports to India from 1850 to 1900. Moretti himself does not focus his work in the digital medium, but his insistence that literary texts can be productively read from a distance as well as up close has provided a critical vocabulary for scholarly projects that use digital visualization tools to wrestle with questions that close reading alone might otherwise be unable to answer.

One such visualization project involves the *Poetess Archive,* a digital archive of poetry from the eighteenth and nineteenth centuries belonging to what project director Laura Mandell refers to as "the 'poetess tradition,' the extraordinarily popular, but much criticized, flowery poetry written in Britain and America between 1750 and 1900."[7] Scholars have known for years that a massive amount of poetry was written and published during this period—a period that Mandell and her collaborators refer to as a "bull market" for poetry. Yet, given the tendency of literary scholarship to focus on a few exceptional poets rather than on an entire poetic scene, the landscape of the poetess tradition has yet to be sufficiently charted. In an effort to fill this gap, Mandell and her collaborators have proposed a visualization tool that will enable visitors to the *Poetess Archive* to import data gleaned from thousands of poetic documents into a program that "will allow users to try out various hypotheses about poetry production during the period," including topics "from metrical forms to semantics, publication venue to graphics on the page, images, book boards, slipcases, etc."[8] A scholar using this tool could generate a list of poems that share similar criteria—for example, poems published in periodicals where illustrations were used to accompany the poetry—and then have the information from this list plotted on a graph with coordinates for, say, date and place of publication. The visualization would then cluster together similar texts into patterns that might not otherwise be apparent, and these resulting patterns would in turn lead to hypotheses about the poetry of the period. Would poems by William Wordsworth, for example, appear anywhere near those of Letitia Elizabeth Landon on such a graph? If so, how might that encourage a scholar to rethink the relationship between High Romanticism and the popular poetry of the poetess tradition?

As the majority of scholars in the digital humanities concur, such visualizations are intended neither to stand as definitive interpretations of literary texts nor to provide direct answers to research questions. Rather, the goal in visualizing data from a literary text (or body of texts) is to spark inquiry. While we might be tempted to think of charts and graphs as the final piece of evidence to definitively nail down an argument (as the talking heads on a cable news show, for example, may use polling data to make claims about the electorate), these scholars in the digital humanities have encouraged us to see visualization tools as a component in a larger interpretative process. Johanna Drucker has referred to this paradigm shift as "a methodological

Given that the entire profession of information visualization has grown up around quantifiable data, it comes as no surprise that literature scholars have been reluctant to turn to graphical representation as a methodology for interpreting literary texts. Literature scholars tend to value close reading—the subtlety of word choice, the nuance of phrasing—over the broad brushstrokes of information visualization. In many ways, the two fields seem to be at a methodological impasse: the virtue of information visualization is that it can make complex data sets more accessible than they might otherwise be, whereas literary close readings often reveal that apparently straightforward texts are more complex than they might otherwise seem. Information visualization seems better suited to analyzing the ups and downs of marketing trends or the changing patterns of crime in a big city than to interpreting the language of literary texts. Nevertheless, scholars working in the digital humanities have found ways to use the tools of information visualization to supplement traditional close readings of literary texts. Instead of parsing out the nuance of individual words and phrases, these scholars have used digital technology to search for patterns and to trace broad outlines, either in a single text or across a body of related texts.

A number of these scholars have cited Franco Moretti's concept of "distant reading" in an effort to differentiate between the traditional practice of close reading and the new ways that digital technologies are allowing literature scholars to read texts. Moretti has argued that close reading of individual texts is not the best way to keep track of the thousands of texts that make up literary history. Rather, he counsels scholars to step back and look at the broad patterns that emerge when you consider a wide swath of texts. He writes that "instead of concrete, individual works" serving as the building blocks of literary history, large-scale patterns of publication and reception provide "a sharper sense of [the] overall interconnection" of texts.[6] Moretti's 2005 book *Graphs, Maps, Trees: Abstract Models for a Literary History* is filled with information graphics that detail, for example, the rise of the British novel from 1700 to 1840 and the number of European book imports to India from 1850 to 1900. Moretti himself does not focus his work in the digital medium, but his insistence that literary texts can be productively read from a distance as well as up close has provided a critical vocabulary for scholarly projects that use digital visualization tools to wrestle with questions that close reading alone might otherwise be unable to answer.

One such visualization project involves the *Poetess Archive,* a digital archive of poetry from the eighteenth and nineteenth centuries belonging to what project director Laura Mandell refers to as "the 'poetess tradition,' the extraordinarily popular, but much criticized, flowery poetry written in Britain and America between 1750 and 1900."[7] Scholars have known for years that a massive amount of poetry was written and published during this period—a period that Mandell and her collaborators refer to as a "bull market" for poetry. Yet, given the tendency of literary scholarship to focus on a few exceptional poets rather than on an entire poetic scene, the landscape of the poetess tradition has yet to be sufficiently charted. In an effort to fill this gap, Mandell and her collaborators have proposed a visualization tool that will enable visitors to the *Poetess Archive* to import data gleaned from thousands of poetic documents into a program that "will allow users to try out various hypotheses about poetry production during the period," including topics "from metrical forms to semantics, publication venue to graphics on the page, images, book boards, slipcases, etc."[8] A scholar using this tool could generate a list of poems that share similar criteria—for example, poems published in periodicals where illustrations were used to accompany the poetry—and then have the information from this list plotted on a graph with coordinates for, say, date and place of publication. The visualization would then cluster together similar texts into patterns that might not otherwise be apparent, and these resulting patterns would in turn lead to hypotheses about the poetry of the period. Would poems by William Wordsworth, for example, appear anywhere near those of Letitia Elizabeth Landon on such a graph? If so, how might that encourage a scholar to rethink the relationship between High Romanticism and the popular poetry of the poetess tradition?

As the majority of scholars in the digital humanities concur, such visualizations are intended neither to stand as definitive interpretations of literary texts nor to provide direct answers to research questions. Rather, the goal in visualizing data from a literary text (or body of texts) is to spark inquiry. While we might be tempted to think of charts and graphs as the final piece of evidence to definitively nail down an argument (as the talking heads on a cable news show, for example, may use polling data to make claims about the electorate), these scholars in the digital humanities have encouraged us to see visualization tools as a component in a larger interpretative process. Johanna Drucker has referred to this paradigm shift as "a methodological reversal which makes visualization a procedure rather than a product and integrates interpretation into digitization in a concrete way."[9] Other scholars, such as those involved in the Nora and MONK projects, have similarly described visualization and related technologies as "instruments for provoking interpretation"—that is, tools that can provoke or inspire inquiry rather than merely answer a specific question—and have posited that a central goal of digital visualization should be, in Matthew Kirschenbaum's words, to "make visualizations function as interfaces in an iterative process that allows [scholars] to explore and tinker."[10] While data visualization may present itself as a scholarly problem-solving tool, these scholars have encouraged us to see visualization as a problem-*generating* tool.

To say that digital visualizations can "provoke" interpretative possibilities is to fulfill, in many ways, the challenge that Jerome McGann laid out for digital literary studies almost a decade ago. "The general field of humanities education and scholarship will not take the use of digital scholarship seriously," McGann wrote, "until one demonstrates how its tools improve the ways we explore and explain aesthetic works—until, that is, they expand our interpretational procedures."[11] The possibility that digital visualization will allow scholars to read and interpret texts differently—by reading them from a distance, for example, rather than up close—is a project that is still very much in its infancy. Nevertheless, there are early indications that visualization tools can help to produce revolutionary interpretations of literary texts. One recent example is Tanya Clement's use of a suite of digital tools developed under the auspices of the MONK project to distant-read Gertrude Stein's 1925 novel, *The Making of Americans.*[12] Twentieth-century critics of Stein's infamously difficult novel have had a love/hate relationship with the text, either dismissing it as "a disaster" whose "tireless and inert repetitiveness . . . amounts in the end to linguistic murder" or praising it as "a postmodern exercise in incomprehensibility that in itself poses a comment on the modernist desire for identity and truth" (Clement, 362). By distant reading *The Making of Americans* with the aid of textual analytics and digital visualization, however, Clement has made the compelling case that Stein's novel is, contrary to the critical commonplaces of the past century, "intricately and purposefully structured" (363).

Given that *The Making of Americans* eschews traditional narrative for a series of oft-repeated words and phrases that Stein seems to sprinkle at random throughout the more than 900 pages of the novel (as Clement notes, "there are 517,207 total words [in the novel] and only 5,329 unique words"), close reading the text has proven to be a frustrating experiences for critics (362). Clement's methodology, in contrast, is to visualize the most commonly repeated words and phrases of the novel—using the FeatureLens software developed in association with MONK as well as more traditional two- and three-dimensional scatter plots—and then to find in these visualizations evidence that Stein had structured her novel according to identifiable patterns of linguistic repetition.[13] Amid "the chaos of the more frequent repetitions," Clement argues, this difficult novel has a deep structure that "readers may have missed with close reading" (363). This structure, she contends, shows that *The Making of Americans* is neither a postmodern exercise in the process of meaning making nor a "disastrous" application of Stein's experimental poetics to the novel form but, instead, a deeply philosophical reflection on the life of an American family.

Aside from the contribution that Clement has made to scholarship on Stein's monumental novel, she has also offered some valuable insight into the challenges of working with digital literary archives. Clement observes that "the particular reading difficulties engendered by the complicated patterns of repetition in *The Making of Americans* mirror those a reader might face attempting to read a large collection of like texts all at once without getting lost" (361). This experience of getting lost among a large collection of texts should resonate with anyone whose initial kid-in-a-candy-store feeling at beginning to work with a rich digital archive of literary texts turned into a deer-in-the-headlights feeling at the prospect of making sense of so vast a repository of information. Matthew Kirschenbaum has noted that "literary scholars . . . traditionally do not contend with very large amounts of data in their research" ("Poetry"). Now that digital literary archives have made it possible for more scholars than ever to access such "very large amounts of data," it has become imperative that we reflect on the ways that we will have to work differently—and even *read* differently—given our access to this expanding body of textual data. Reading distantly is one option; reading spatially is another.

Spatial Reading

The field of information visualization was born, as Usama Fayyad and Georges G. Grinstein write, from "the explosive generation of massive data sets and our need to extract the data's inherent information."[14] A comparable "explosion" is taking place in the study of American literature as digital archives are becoming an increasingly important part of our teaching and scholarship. While digital visualization tools are poised to deal with a similar set of issues as those faced by our colleagues in the sciences and social sciences, some of the assumptions about reading expressed in the scholarship on information visualization tend not to sit well with scholars and teachers of literature. We might balk at many of these assumptions, but as professional readers—which, among other things, is what literature scholars are—it behooves us to be involved in the conversations that are taking place about the fate of reading in an era of digital visualization.

For example, a decade or so ago, a group of research scientists in the field of information visualization claimed that the realities of the digital age necessitated a fundamental change in the way that people read. "Modern information technologies," they argued, "have made so much text available that it overwhelms the traditional reading methods of inspection, sift and synthesis."[15] As a way to deal with this "overwhelming" proliferation of texts, they proposed that computer-generated visualizations of text patterns would be able to "reduce [readers'] mental workload" by extracting the valuable information from a text so that readers would not "hav[e] to read it in the manner that text normally requires" ("Visualizing," 442). Most teachers and scholars of literature—and I include myself in this group—have an immediate knee-jerk reaction to statements such as these. Given that we spend so much of our professional lives encouraging people to read more rather than less and to read slowly and carefully rather than in a quick and cursory manner, the prospect of developing technological means for reducing readers' "mental workload" and thereby freeing them from an intellectual process of "inspection, sift and synthesis" seems anathema to what we think the experience of reading should be. Similarly, when computer scientists and graphic designers argue that "human intuition can be more of a hindrance than a helpful factor" for finding the pertinent information in large amounts of text ("Introduction," 2) or when they prophesy that "the limitations of an Information Age will not be set by the speed with which a human mind can read" ("Visualizing," 449), it is hard not to cringe.

Granted, the kinds of text that scholars in the field of information visualization have traditionally been concerned with are not nuanced works of imaginative literature but information-rich documents filled with medical, scientific, and other types of quantifiable data. The fact that the information visualization community has already expressed interest in how to visualize literary texts, however, makes it all the more important for literature scholars to join this conversation. By joining it, not only will we be able to share what we have learned about the distinct properties of literary texts, but we will also, in the spirit of interdisciplinarity (if not humility), be in a position to learn something about the challenges involved in reading large amounts of texts. Specifically, literature scholars could do with a crash course in the cognition of reading, a topic that many scholars of information visualization have spent a good bit of time thinking about.

By and large, literature scholars are not only people of the book and people of the word but people of the-typeface-that-does-not-call-attention-to-itself. We tend to assume that knowledge is transmitted through those supposedly transparent carriers of thought: printed words. Many scholars in the field of information visualization, in comparison, have taken to studying the cognitive and perceptual dynamics that shape the reading process, pondering, for example, the ways in which visual stimuli such as shape, color, and texture affect the brain's ability to process information. Their effort to create visual abstractions of textual patterns is motivated not only by a desire to use technology to speed up the reading process but also by an eagerness to learn more about the workings of the human mind.

One of the main concepts driving the recent research on reading, cognition, and digital visualization is the notion that the mind is just as capable (if not more so) of extracting meaning from shapes and patterns as it is at processing written language. As one group of scholars has written, "Humans are quite adept at perceptual visual cues and recognizing subtle shape differences. In fact, it has been shown that humans can distinguish shape during the pre-attentive psychophysical process." Because, they continue, "humans are pre-wired for understanding and visualizing shape," digital tools that transform textual patterns into visual shapes will assist readers in "harnessing these skills of shape perception."[16] The idea that there is a preattentive information process, or (as another group puts it) "a preconscious visual form for information" whereby the mind intuitively recognizes and comprehends patterns of meaning, has led a number of scholars to speculate that digital visualizations will accelerate the reading process by allowing readers to access that portion of the mind that processes information spatially rather than sequentially ("Visualizing," 445).[17]

Along these lines, one group of research scientists has argued that "the bottleneck in the human processing and understanding of information in large amounts of text can be overcome if the text is spatialized in a manner that takes advantage of common powers of perception" ("Visualizing," 443). The motivation to create digital tools for "transforming the text information to a spatial representation which may then be accessed and explored by visual processes" emerges from a desire to privilege readers' capacity for spatial perception over their usual habit of sequential reading. In so doing, the thinking goes, readers will then be able to escape "the rather slow serial process of mentally encoding a text document" and "instead use their primarily preattentive, parallel processing powers of visual perception" ("Visualizing," 442). While literature scholars tend to assume that reading is a necessarily sequential act—for us, reading usually means following a string of words from beginning to end—a number of scholars and professionals in the field of information visualization have attempted to represent the meaningful patterns in a corpus of texts as "concept shapes" whose meaning can be quickly apprehended by the brain's natural propensity for spatial recognition.

Creating "concept shapes" out of texts is similar to graphically representing data patterns with more conventional visualizations, such as scatter plots. In a scatter plot, data are charted onto a graph so that an analyst can observe the patterns that emerge as data points cluster together relative to the axes *x*, *y*, and *z* that define the boundaries of the graph. Similarly, in an effort to help readers of large textual corpora "better understand document content and relationships," one group of scholars has devised a method for representing texts as semispherical objects in a virtually rendered three-dimensional space ("Shape," 1). When texts in a document corpus demonstrate patterns of similarity (based on such factors as, say, common word usage), these spherical objects blend together to create a variety of quasi-organic shapes referred to as "blobby models, meatballs, or soft objects" ("Shape," 2). As part of their experiment in visualizing text patterns as blobs of virtual goo, these scholars also took a crack at literary analysis. Figure 1, which I have taken from their published findings, "shows a detailed example of three documents. The two that are clustered together are Shakespeare's plays *Richard II* and *Richard III* while the solitary document within its own cluster is a document on information visualization (two vastly different concepts from vastly different ages)" ("Shape," 7). It is both thrilling and, to be honest, a little disturbing to watch Shakespeare's plays digitally morphed into a shape resembling nothing so much as a mutated chicken embryo. Nevertheless, I am reminded that such experiments in text visualization are motivated by the hope that physical shapes—more so than, say, the pinpoints on a scatter plot—will not only be able to trigger the mind's capacity for spatial recognition but will also allow readers to quickly and intuitively identify the patterns that might otherwise be overlooked when reading a large body of texts.

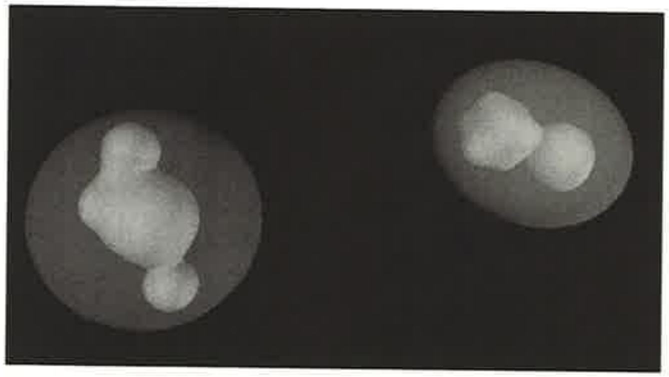

Fig. 1. Information visualization cluster and Shakespeare cluster (*Richard II*, *Richard III*). (From Randall M. Rohrer, David S. Ebert, and John L. Sibert, "The Shape of Shakespeare: Visualizing Text Using Implicit Surfaces," *Proceedings of the 1998 IEEE Symposium on Information Visualization* [Washington, DC: IEEE Computer Society, 1998], 3, http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00729568.)

A related example comes from another group of research scientists, who, following a similar line of inquiry, have postulated that "spatializing text content for enhanced visual browsing and analysis" functions best when readers are given "an interaction with text that more nearly resembles perception and action with the natural world" ("Visualizing," 442). The resulting visualization tool that they have devised uses clusters of data points (again, similar to those in a scatter plot) as the basis for what they refer to as "galaxy visualizations." A galaxy visualization projects points of light, which represent information gleaned from a group of text documents, onto a black background, in a manner that is designed to "recapitulate experiences of viewing the night sky" ("Visualizing," 448–49). When meaningful patterns appear in the data, clustered points of light appear as constellations within the larger "galaxy." The conceit behind this visualization is that the same human capacity for finding meaning in the stars—the same capacity, that is, that anciently populated the night sky with gods and heroes—continues to function on a computer screen. Along the same lines, this group has also created a method for viewing data points as elements in a textured, three-dimensional wave—which they describe as a "visual metaphor" for "traversing landscapes"—that not only presents readers with an image reminiscent of the geographical contours of the earth's surface but also uses these data-driven peaks and valleys to evoke humans' innate ability to spatially process their physical environment ("Visualizing," 448–49). As with their galaxy visualization, the implication of the landscape visualization is that the "primitive" cognition of hunter-gatherers buried deep in the evolutionary recesses of the mind can be reawakened in the digital age as a way to find meaningful patterns in large bodies of textual data.

Despite the neo-Romantic organicism that permeates these attempts to represent text patterns as shapes from the natural world, there is also a tacit acknowledgment by these same practitioners of information visualization that there is something fundamentally *un*natural about efforts to render words as images. Computer scientists and graphic designers often refer to text visualizations as attempts to "visualize the nonvisual," a formulation that suggests the irony, if not the futility, of making visual perception an integral part of the reading process. "Since visualizing text requires mapping the abstract to the physical," writes the group of scholars behind the quasi-organic "blobby" text models, the primary challenge facing any project in text visualization lies in creating an "interface for providing [a] layer of abstraction" between the original text and the resulting visual image ("Shape," 2). While blobby models and virtual landscapes represent fascinating attempts at designing an interface between word and image, some of the most promising text visualization projects of recent years take as their starting point the idea that words *are* images and that the search for a "layer of abstraction" between word and image has created a false dichotomy between reading and seeing.[18] By experimenting with such elementary bibliographic signifiers as font size and the arrangement of words on the page (or screen), a new generation of text visualization projects have suggested possibilities for spatial reading that treat text as both words to be read and shapes to be viewed.

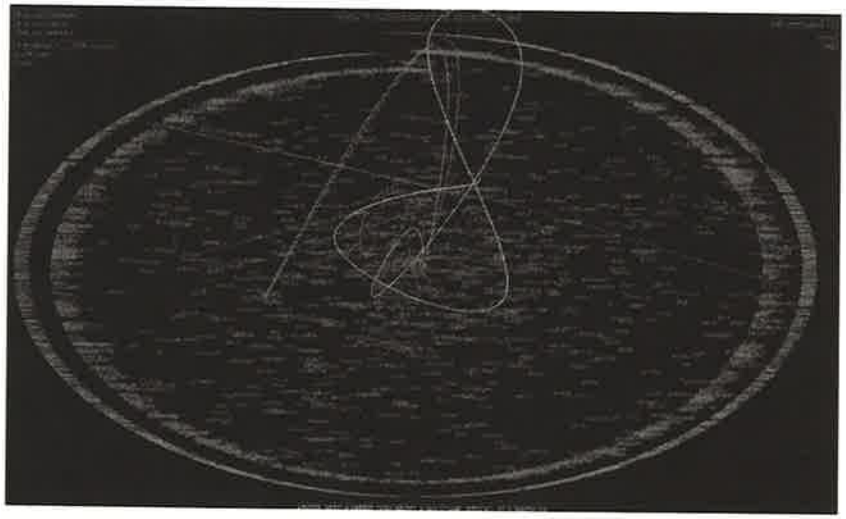

Fig. 2. W. Bradford Paley's TextArc rendition of Lewis Carroll's novel *Alice in Wonderland,* http://www.textarc.org/images/alice1.gif

One example of a text visualization process that retains the text *as* the visualization is W. Bradford Paley's 2002 TextArc project. Paley recites many of the same goals for visualizing text as do the scholars and professionals behind other projects. Writing, for example, that he wants "to help people discover patterns and concepts in any text by leveraging a powerful, underused resource: human visual processing," he claims that his project not only "taps into our pre-attentive ability to scan" for visual patterns of meaning but also facilitates a reader's "intuition [to] extract meaning from an unread text."[19] Despite these similarities, the visualizations produced by TextArc minimize (if not collapse) the need for a "layer of abstraction" between text and image found in other projects. A TextArc visualization, such as the one for Lewis Carroll's *Alice in Wonderland* in figure 2, is an image comprised entirely of words: the ellipse that frames the screen is a word-for-word reproduction of the complete text (in a one-pixel font), and the amorphous cloud that fills out the center of the ellipse is a color-coded array of the text's most commonly used words (oft-repeated words glow brighter than words that do not occur as frequently; and words that appear throughout the text migrate to the center of the cloud, while words that are specific to a given section of the text tend toward its peripheries). The visualization is also interactive. As the cursor floats over an individual word in the cloud, for example, rays of light connect that word to its occurrences throughout the text ellipse; and, if requested, a traditional concordance or keyword-in-context index can be generated alongside the visualization.

TextArc is, among other things, an experiment in spatial reading that is grounded in the belief that reading and seeing are complementary processes. Paley describes TextArc as a "balancing act" between reading and seeing, explaining that as readers experience the text visualization, "the eye and mind scan for ideas, then follow the ideas down to where and how they appear in the text" ("TextArc"). Such forms of digital textuality that blur the line between text and image are still very much in their infancy, and the interdisciplinary research into how, precisely, the eye and the mind process information in this format has yet to be fully conducted. While the academic community awaits the outcomes of this research, ambitious graphic designers and computer programmers have already begun populating the World Wide Web with text visualization tools that allow anyone with Internet access to upload the text of their choice and create a word cloud similar to the numinous field of text at the center of a TextArc. Around the middle of the 2000s, popular photo- and file-sharing sites began using tag clouds to indicate which descriptors (or "tags") were most frequently used to categorize files and photos, with larger-font tags indicating a higher frequency of usage than smaller-font tags. Since then, tag and word clouds have become ubiquitous on the Web.[20] Word clouds have proven to be quite popular with Internet users, both for their playful aesthetic quality and for their practical ability to visually identify the patterns of meaning in large and potentially unwieldy texts.[21]

Literature scholars have yet to fully theorize the ways in which spatially reading the text in a word cloud can lead to new and exciting interpretative possibilities.[22] As an informal experiment in spatial reading, I found a word cloud of Walt Whitman's "Song of Myself" on Wordle.net, a popular word-cloud generator, and attempted to compare my past experiences reading Whitman's monumental 1,300-line poem with the experience of reading it spatially as a digital cloud (see fig. 3). Reading/viewing "Song of Myself" as a cloud of words immediately refamiliarized me with a poem I have read many times before, but it also defamiliarized a poem that I thought I knew so well. I was not surprised at all to see words like *love, earth, see,* and *know* jump out of the cloud, but I was shocked to see the words *shall* and

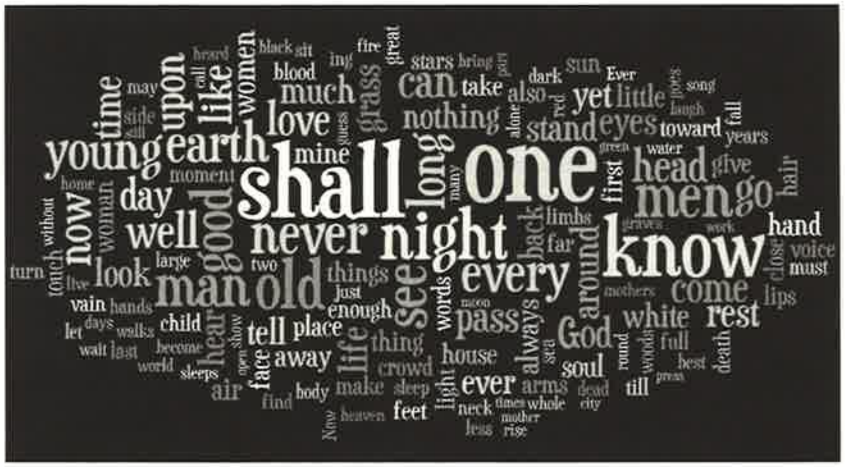

Fig. 3. Word cloud of Walt Whitman's "Song of Myself." (From Wordle.net, http://www.wordle.net/gallery/wrdl/180308/Song_of_Myself_-_Walt_Whitman.)

one emerge with such prominence. I tend to associate the word *shall* with the proscriptive language of the Bible, with its commandments of "Thou shall not" and "Thou shall." Whitman has always struck me as the poet of laissez-faire, content to observe rather than prescribe. But the word cloud reminded me that he is also a poet of the future, of possibility, of action—the poet, that is, of "shall." The word *one* had a similarly defamiliarizing effect on me. I had always thought of Whitman as a poet of diversity and expanse, of the many rather than the one. But reading "Song of Myself" in this format reminded me of the centripetal as well as centrifugal pull in Whitman's poetry, of his tendency to collapse all experience into the unity of the self. I often tell my students that Walt Whitman and Emily Dickinson teach us to read in very different ways: Dickinson requires us to drill down into the meaning of specific words if we are to make sense of the larger poem, whereas Whitman requires us to step back and get a sense of the entire landscape of the poem in order to grasp its meaning. I found, in this entirely unsystematic and wholly impressionistic exercise in spatial reading, that the word cloud of "Song of Myself" rekindled that sentiment for me in exciting and thought-provoking ways.[23]

At least two other scholars working in the digital humanities—Lisa Spiro and Sara Steger—have made similar attempts to read nineteenth-century literature as a cloud of digital text. Both Spiro and Steger have used word clouds generated with Wordle, along with other text analysis tools, to rethink the language of literary sentimentalism. Sentimentalism is a broad and often oversimplified term in literary studies, and Spiro and Steger make welcome additions to scholarship from the past two decades that has challenged preconceived notions about sentimental literature.[24] Steger's project involved running nearly four thousand mid-Victorian novels through digital text analysis tools available through MONK, sifting out the words and phrases most often identified as sentimental, and then using Wordle to visualize the patterns that emerged. Steger's preliminary findings were hardly controversial: she found, for example, that deathbed scenes in sentimental novels employ "vocabulary [that] emphasizes intimate relationships—'mamma,' 'papa,' 'darling,' and 'child.'" But Steger's greatest insights come not from the moments where the visualization highlights the most commonly appearing words but, rather, from those where it shows her "that which is absent." "What the word cloud does not include is almost as informative as what it does," she writes. "One of the most under-represented words is 'holy,' and it is followed by 'church,' 'saint,' 'faith,' 'believe' and 'truth.' It seems the Victorian deathbed scene is more concerned with relationships . . . than with personal convictions and declarations of faith."[25] For many readers, literary sentimentalism is inextricably connected to the larger religious worldview from which it is presumed to have emerged. Steger has found, in contrast, a much more complex relationship between sentimental discourse and nineteenth-century religious language.

Steger's use of word clouds to spatially read a large body of texts involves an interesting back-and-forth between close and distant reading: at one moment, her wide-angle perspective charts the broad contours of sentimental language across a vast array of texts, while at other moments, her intense focus on specific words feels like close reading on a microscopic scale. Lisa Spiro makes a similar move in her word-cloud analysis of the sentimental language in Donald Grant Mitchell's *Reveries of a Bachelor* (1850) and Herman Melville's *Pierre* (1852), a text that she argues is a "dark parody" of Mitchell's *Reveries*. Spiro's concern is with Melville's appropriation and transformation of sentimental language used in mainstream texts such as Mitchell's, and she uses word clouds to compare and contrast word frequency and usage between the two. Spiro comes to a number of thought-provoking conclusions in the course of her analysis—among them, that Melville often uses the same words as those employed in more traditional

sentimental texts but derives from them a "different resonance," taking, as she puts it, "some of the ingredients of sentimental literature and mak[ing] something entirely different with them." But what most interests me about her methodology is the similar back-and-forth between close and distant reading that Steger also employs in her analysis of mid-Victorian literature (and, for that matter, that I use in my own informal spatial reading of the "Song of Myself" word cloud). Spiro writes that the spatialized text in the world cloud provided her with the "initial impression" that inspired her analysis of the two texts, but she then goes on to note that what made the most significant impact on her analysis was not the shape of the cloud itself but the quantitative values that determined which words would pop out of the cloud as larger and which would recede into the background as smaller. "Ultimately," she writes, "I trusted the concreteness and specificity of numbers more than the more impressionistic imagery provided by the word cloud." Despite granting this authority to quantitative analysis, however, she is quick to caveat that "the word cloud opened up my eyes so that I could see the stats more meaningfully."[26]

There seems to be a push-and-pull involved in spatially reading a word cloud, as Spiro, Steger, and I all found ourselves alternating between observing the big picture and honing in on specific words. Spatial reading is a curious hybrid of close and distant reading, it seems, requiring both impressionistic reactions and quantitative analysis. This push toward the quantitative serves as a reminder that digital visualization often requires that we reduce language—that plastic, ambiguous, free-form media we scholars of literature love to play in—to the stable, albeit more dour, realm of numbers. By the same token, word clouds promise to keep the tension between words and numbers—not to mention images—at play in provocative and exciting ways. Whether or not the methods of reading and interpretative discovery provoked by word clouds (or by any digital visualization tool, for that matter) will become a part of our critical practice as scholars and teachers of literature remains to be seen. Again, such technologies are still in their infancy, but it bears noting that these infant technologies are growing up alongside our own still-young archives of digitized text. The forces of the digital era are rethinking the ways that we read at the same time that American literature scholars are rethinking the ways that we archive large bodies of texts. It would benefit both parties to pay closer attention to what the other is doing.

Notes

1. Samantha Power, "The Short Tail," *Time,* 2 June 2008, 34.

2. The digital archive *Uncle Tom's Cabin & American Culture* (http://www.iath.virginia.edu/utc/) uses the terms *search mode* and *browse mode* explicitly, while other archives implicitly structure the user experience in this way. There are notable exceptions, of course. The *Willa Cather Archive,* for example, has integrated TokenX, a text analysis and visualization tool, into the archive itself (see http://cather.unl.edu/tokenx.intro.html), as has the *Walt Whitman Archive* (see http://www.whitmanarchive.org/resources/tools/index.html).

3. Stephen Ramsay, "Databases," in *A Companion to Digital Humanities,* ed. Susan Schreibman, Ray Siemens, and John Unsworth (Malden, MA : Blackwell, 2004), 177–97, at 195. See the discussion of the database structure of archives—as both a literal and metaphorical structure—in *PMLA* 122, no. 5 (October 2007): Ed Folsom, "Database as Genre: The Epic Transformation of Archives," 1571–79; Jerome McGann's response, "Database, Interface, and Archival Fever," 1588–92; and Folsom's subsequent rejoinder, "Reply," 1608–12.

4. The literature on digital visualization has increased in both quantity and quality in recent years. For two particularly useful essays that overview innovation in the field, see John Risch et al., "Text Visualization for Visual Text Analytics," in *Visual Data Mining: Theory, Techniques, and Tools for Visual Analytics,* ed. Simeon Simoff, Michael Böhlen, and Arturas Mazeika (Berlin, Heidelberg, and New York: Springer, 2008), 154–71; Martyn Jessop, "Digital Visualization as a Scholarly Activity," *Literary and Linguistic Computing* 23, no. 3 (September 2008), 281–93. The essay by Risch et al. provides an overview of the technological and methodological innovations in text visualization in recent years, while Jessop's essay focuses on the applicability of digital visualization to humanities research.

5. In an effort to focus on the methodological and theoretical assumptions behind digital text visualization tools and how those assumptions affect the ways in which literature scholars tend to think about the reading process, I have dedicated the majority of this essay to a critical examination of these assumptions rather than to a survey of the text visualization projects that have been developed for the digital medium. Lisa Spiro has helpfully cataloged many of the most cutting-edge text visualization tools at http://www.diigo.com/user/lspiro/text_visualization.

6. Franco Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History* (New York: Verso, 2005), 1. Moretti first put forward this idea in the essay "Conjectures on World Literature," *New Left Review* 1 (January–February 2000), 54–68. Moretti was the keynote speaker at the 2007 meeting of the Association for Computers and the Humanities, where he spoke on the topic of distant reading and digital scholarship.

7. "Introduction to the Poetess Archive," *Poetess Archive,* http://unixgen.muohio.edu/~poetess/about/index.html (accessed 28 October 2008).

8. "Coming Soon: Visualization Tool for Poetic Elements, 1750–1850," *Poetess Archive,* http://unixgen.muohio.edu/~poetess/vmodel/vmodel.html (accessed 28 October 2008).

9. Johanna Drucker and Bethany Nowviskie, "Speculative Computing: Aesthetic Provocations in Humanities Computing," in Schreibman, Siemens, and Unsworth, *Companion to Digital Humanities,* 431–47, at 442.

10. Matthew Kirschenbaum, "Poetry, Patterns, and Provocation: The nora Project," *The Valve: A Literary Organ,* 12 January 2006, http://www.thevalve.org/go/valve/article/poetry_patterns_and_provocation_the_nora_project/ (accessed 10 August 2008) (hereafter cited in text as "Poetry"). For other research by scholars affiliated with the Nora and MONK projects, see Catherine Plaisant et al., "Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces," in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (New York: ACM Press, 2006), 141–50, at 141.

11. Jerome McGann, *Radiant Textuality: Literature after the World Wide Web* (London: Palgrave Macmillan, 2001), xii.

12. Tanya Clement, "'A thing not beginning and not ending': Using Digital Tools to Distant-Read Gertrude Stein's *The Making of Americans,*" *Literary and Linguistic Computing* 23, no. 3 (September 2008): 361–81 (hereafter cited in text as Clement).

13. For a description of the FeatureLens software, see Anthony Don et al., "Discovering Interesting Usage Patterns in Text Collections: Integrating Text Mining with Visualization," in *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management* (New York: ACM Press, 2007), 213–22.

14. Usama Fayyad and Georges G. Grinstein, introduction to *Information Visualization in Data Mining and Knowledge Discovery,* ed. Usama Fayyad, Georges G. Grinstein, and Andreas Wierse (San Francisco: Morgan Kaufmann, 2001), 1–17, at 1 (hereafter cited in text as "Introduction").

15. James A. Wise et al., "Visualizing the Non-visual: Spatial Analysis and Interaction with Information from Text Documents," in *Readings in Information Visualization: Using Vision to Think,* ed. Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman (San Francisco: Morgan Kaufmann, 1999), 442–49, at 442 (hereafter cited in text as "Visualizing").

16. Randall M. Rohrer, David S. Ebert, and John L. Sibert, "The Shape of Shakespeare: Visualizing Text Using Implicit Surfaces," in *Proceedings of the 1998 IEEE Symposium on Information Visualization* (Washington, DC: IEEE Computer Society, 1998), 3, http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00729568 (accessed 20 August 2008) (hereafter cited in text as "Shape").

17. See also Andrew J. Parker et al., "The Analysis of 3D Shape: Psychophysical Principles and Neural Mechanisms," in *Understanding Vision: An Interdisciplinary Perspective,* ed. Glyn Humphreys (Malden, MA: Blackwell, 1992).

18. For more on the word/image and reading/seeing dichotomy, see Matthew G. Kirschenbaum, "The Word as Image in an Age of Digital Reproduction," in *Eloquent*

Images: Word and Image in the Age of New Media (Cambridge, MA: MIT Press, 2003), 137–56; Johanna Drucker and Charles Bernstein, *Figuring the Word: Essays on Books, Writing, and Visual Poetics* (New York: Granary Books, 1998); Berjouli Bowler, *The Word as Image* (London: Studio Vista, 1970).

19. W. Bradford Paley, "TextArc: Revealing Word Associations, Distribution, and Frequency," 2002, http://www.textarc.org/TextArcOverview.pdf (accessed 20 August 2008) (hereafter cited as "TextArc"). The entire project is available online at http://www.textarc.org/. It bears noting that, as do other text visualization project directors, Paley relies on a set of organic metaphors to describe the cognitive processes involved in reading a TextArc: "A botanist learns visual strategies for distinguishing the type and health of a plant; likewise people looking at TextArcs have begun to develop visual strategies that help extract structural features in texts" ("TextArc").

20. For a scholarly overview of the phenomenon of tag clouds, see Martin Halvey and Mark T. Keane, "An Assessment of Tag Presentation Techniques," *Proceedings of the 16th International Conference on the World Wide Web* (New York: ACM Press, 2007), 1313–14; A. W. Rivadeneira et al., "Getting Our Head in the Clouds: Toward Evaluation Studies of Tagclouds," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York: ACM Press, 2007), 995–98.

21. As one Web developer recently commented, "Whether it's a campaign speech by a presidential contender, or a 300-page bestselling novel, large bodies of text are among the most requested topics for condensing into an infographic. The purpose can vary from highlighting specific relations to contrasting points or use of language, but all [such visualization tools] focus on distilling a volume of text down to a visualization" (Tim Showers, "Visualization Strategies: Text and Documents" August 2008, http://www.timshowers.com/2008/08/visualization-strategies-text-documents/ [accessed 31 October 2008]).

22. For an attempt to use word clouds for literary analysis, see Lisa Spiro, "Using Text Analysis Tools for Comparison: Mole & Chocolate Cake," 22 June 2008, http://digitalscholarship.wordpress.com/2008/06/22/using-text-analysis-tools-for-comparison-mole-chocolate-cake/ (accessed 10 August 2008).

23. Not only is reading a word cloud an admittedly subjective experience, but *creating* a word cloud is also highly subjective. The world cloud of "Song of Myself" I found on Wordle.net, for example, was designed such that commonly occurring words (e.g., *I, of, the, a, an*) were excluded. The choice of color, size, font, and arrangement of these words also no doubt influenced how I spatially read the poem. The fact that many different word-cloud versions of this same poem could be (and, indeed, have already been) created on such sites as Wordle.net does not, I believe, discredit the new interpretative possibilities that word clouds have to offer; it instead demands that readers of such clouds be self-reflexive as they read such texts spatially. I would argue that such human-computer interaction should be seen not as a limitation of digital visualization but, rather, as a productive site of possibility. For more on human-computer interaction (HCI) and its relation to digital visualization, see Ben Shneiderman, "Inventing Discovery Tools: Combining Information Visualization with Data Mining," in *The Craft of Information Visualization: Readings and Reflections*, ed. Benjamin B. Bederson and Ben Shneiderman (New York: Morgan Kaufmann, 2003), 379–85.

24. For representative texts in this scholarship on sentimental literature, see Jane Tompkins, *Sensational Designs: The Cultural Work of American Fiction, 1790–1860* (New York: Oxford University Press, 1985); Joanne Dobson, "Reclaiming Sentimental Literature," *American Literature* 69 (June 1997): 263–88; and Elizabeth Maddock Dillon, "Sentimental Aesthetics," *American Literature* 76 (September 2004): 495–523.

25. Steger's work is described in a collaborative presentation by Tanya Clement, Sara Steger, John Unsworth, and Kirsten Uszkalo titled "How Not to Read a Million Books," presented originally at the Seminar in the History of the Book at Rutgers University, New Brunswick, NJ, 5 March 2009, and available online at http://www3.isrl.illinois.edu/~unsworth/hownot2read.html (accessed 31 August 2009). My thanks to Meredith McGill and Lisa Gitelman for bringing Steger's work to my attention.

26. Spiro, "Using Text Analysis Tools."