

1 REVOLUTION

The digital revolution is far more significant
than the invention of writing or even of printing.

—Douglas Carl Engelbart

An article in the June 23, 2008, issue of *Wired* declared in its headline “Data Deluge Makes the Scientific Method Obsolete” (Anderson 2008). By 2008 computers, with their capacity for number crunching and processing large-scale data sets, had revolutionized the way that scientific research gets done, so much so that the same article declared an end to theorizing in science. With so much data, we could just run the numbers and reach a conclusion. Now slowly and surely, the same elements that have had such an impact on the sciences are revolutionizing the way that research in the humanities gets done. This emerging field we have come to call “digital humanities”—which was for a good many decades not emerging at all but known as “humanities computing”—has a rich history dating back at least to Father Roberto Busa’s concordance work in the 1940s, if not before.* Only recently, however, has this “discipline,” or “community of practice,” or “field of study/theory/methodology,” and so on, entered into the mainstream discourse of the humanities, and it is even more recently that those who “practice” digital humanities (DH) have begun to grapple with the challenges of big data.† Technology has certainly changed some things about the way literary scholars go about their work, but until recently change has been

* Roberto Busa, a Jesuit priest and scholar, is considered by many to be the founding father of humanities computing. He is the author of the *Index Thomisticus*, a lemmatized index of the works of Thomas Aquinas.

† Some have already begun thinking big. In 2008 I served on the inaugural panel reviewing applications for the jointly sponsored National Endowment for the Humanities and National Science Foundation “Digging into Data” grants. The expressed goals of the grant are to promote the development and deployment of innovative research techniques in large-scale data analysis; to foster interdisciplinary collaboration among scholars in

mostly at the level of simple, even anecdotal, search. The humanities computing/digital humanities revolution has now begun, and big data have been a major catalyst. The questions we may now ask were previously inconceivable, and to answer these questions requires a new methodology, a new way of thinking about our object of study.

For whatever reasons, be they practical or theoretical, humanists have tended to resist or avoid computational approaches to the study of literature.* And who could blame them? Until recently, the amount of knowledge that might be gained from a computer-based analysis of a text was generally overwhelmed by the dizzying amount of work involved in preparing (digitizing) and then processing that digital text. Even as digital texts became more readily available, the computational methods for analyzing them remained quite primitive. Word-frequency lists, concordances, and keyword-in-context (KWIC) lists are useful for certain types of analysis, but these staples of the digital humanist's diet hardly satiate the appetite for more. These tools only scratch the surface in terms of the infinite ways we might read, access, and make meaning of text. Revolutions take time; this one is only just beginning, and it is the existence of digital libraries, of large electronic text collections, that is fomenting the revolution. This was a moment that Rosanne Potter predicted back in the digital dark ages of 1988. In an article titled "Literary Criticism and Literary Computing," Potter wrote that "until everything has been encoded, or until encoding is a trivial part of the work, the everyday critic will probably not consider computer treatments of texts" (93). Though not "everything" has been digitized, we have reached a tipping point, an event horizon where enough text and literature have been encoded to both allow and, indeed, force us to ask an entirely new set of questions about literature and the literary record.

the humanities, social sciences, computer sciences, information sciences, and other fields around questions of text and data analysis; to promote international collaboration; and to work with data repositories that hold large digital collections to ensure efficient access to these materials for research. See <http://www.diggingintodata.org/>.

* I suspect that at least a few humanists have been turned off by one or more of the very public failures of computing in the humanities: for example, the Donald Foster Shakespeare kerfuffle.

2 EVIDENCE

Scientists scoff at each other's theories but agree
in basing them on the assumption that evidence,
properly observed and measured, is true.

—Felipe Fernández-Armesto

While still graduate students in the early 1990s, my wife and I invited some friends to share Thanksgiving dinner. One of the friends was, like my wife and me, a graduate student in English. The other, however, was an outsider, a graduate student from geology. The conversation that night ranged over a wine-fueled spectrum of topics, but as three of the four of us were English majors, things eventually came around to literature. There was controversy when we came to discuss the “critical enterprise” and what it means to engage in literary research. The very term *research* was discussed and debated, with the lone scientist in the group suggesting, asserting, that the “methodology” employed by literary scholars was a rather subjective and highly anecdotal one, one that produced little in terms of “verifiable results” if much in the way of unsupportable speculation.

I recall rising to this challenge, asserting that the literary methodology was in essence no different from the scientific one: I argued that scholars of literature (at least scholars of the idealistic kind that I then saw myself becoming), like their counterparts in the sciences, should and do seek to uncover evidence and discover meaning, perhaps even truth. I dug deeper, arguing that literary scholars employ the same methods of investigation as scientists: we form a hypothesis about a literary work and then engage in a process of gathering evidence to test that hypothesis.

After so many years it is only a slightly embarrassing story. Although I am no longer convinced that the methods employed in literary studies are exactly the same as those employed in the sciences, I remain convinced that there are a good many methods worth sharing and that the similarities of methods exist in concrete ways, not simply as analogous practices.

The goal of science, we hope, is to develop the best possible explanation for some phenomenon. This is done via a careful and exhaustive gathering of evi-

dence. We understand that the conclusions drawn are only as good as the evidence gathered, and we hope that the gathering of evidence is done both ethically and completely. If and when new evidence is discovered, prior conclusions may need to be revised or abandoned—such was the case with the Ptolemaic model of a geocentric universe. Science is flexible in this matter of new evidence and is open to the possibility that new methods of investigation will unearth new, and sometimes contradictory, evidence.

Literary studies should strive for a similar goal, even if we persist in a belief that literary interpretation is a matter of opinion. Frankly, some opinions are better than others: better informed, better derived, or just simply better for being more reasonable, more believable. Science has sought to derive conclusions based on evidence, and in the ideal, science is open to new methodologies. Moreover, to the extent possible, science attempts to be exhaustive in the gathering of the evidence and must therefore welcome new modes of exploration, discovery, and analysis. The same might be said of literary scholars, excepting, of course, that the methods employed for the evidence gathering, for the discovery, are rather different. Literary criticism relies heavily on associations as evidence. Even though the notions of evidence are different, it is reasonable to insist that some associations are better than others.

The study of literature relies upon careful observation, the sustained, concentrated reading of text. This, our primary methodology, is “close reading.” Science has a methodological advantage in the use of experimentation. Experimentation offers a method through which competing observations and conclusions may be tested and ruled out. With a few exceptions, there is no obvious corollary to scientific experimentation in literary studies. The conclusions we reach as literary scholars are rarely “testable” in the way that scientific conclusions are testable. And the conclusions we reach as literary scholars are rarely “repeatable” in the way that scientific experiments are repeatable. We are highly invested in interpretations, and it is very difficult to “rule out” an interpretation. That said, as a way of enriching a reader’s experience of a given text, close reading is obviously fruitful; a scholar’s interpretation of a text may help another reader to “see” or observe in the text elements that might have otherwise remained latent. Even a layman’s interpretations may lead another reader to a more profound, more pleasurable understanding of a text. It would be wasteful and futile to debate the value of interpretation, but interpretation is fueled by observation, and as a method of evidence gathering, observation—both in the sciences and in the humanities—is flawed. Despite all their efforts to repress them, researchers will have irrepressible biases. Even scientists will “interpret” their evidence through a lens of subjectivity. Observation is flawed in the same way that generalization from the specific is flawed: the generalization may be good, it may even explain a total population, but the selection of the sample is always something less than

perfect, and so the observed results are likewise imperfect. In the sciences, a great deal of time and energy goes into the proper construction of “representative samples,” but even with good sampling techniques and careful statistical calculations, there remain problems: outliers, exceptions, and so on. Perfection in sampling is just not possible.

Today, however, the ubiquity of data, so-called big data, is changing the sampling game. Indeed, big data are fundamentally altering the way that much science and social science get done. The existence of huge data sets means that many areas of research are no longer dependent upon controlled, artificial experiments or upon observations derived from data sampling. Instead of conducting controlled experiments on samples and then extrapolating from the specific to the general or from the close to the distant, these massive data sets are allowing for investigations at a scale that reaches or approaches a point of being comprehensive. The once inaccessible “population” has become accessible and is fast replacing the random and representative sample.

In literary studies, we have the equivalent of this big data in the form of big libraries. These massive digital-text collections—from vendors such as Chadwyck-Healey, from grassroots organizations such as Project Gutenberg, from nonprofit groups such as the Internet Archive and HathiTrust, and from the elephants in Mountain View, California, and Seattle, Washington*—are changing how literary studies get done. Science has welcomed big data and scaled its methods accordingly. With a huge amount of digital-textual data, we must do the same. Close reading is not only impractical as a means of evidence gathering in the digital library, but big data render it totally inappropriate as a method of studying literary history. This is not to imply that scholars have been wholly unsuccessful in employing close reading to the study of literary history. A careful reader, such as Ian Watt, argues that elements leading to the rise of the novel could be detected and teased out of the writings of Defoe, Richardson, and Fielding. Watt’s study is magnificent; his many observations are reasonable, and there is soundness about them.† He appears correct on a number of points, but he has observed only a small space. What are we to do with the other three to five thousand works of

* That is, Google.com and Amazon.com.

† A similar statement could be made of Erich Auerbach’s *Mimesis*. It is a magnificent bit of close reading. At the same time, Auerbach was acutely aware of the limitations of his methodology. In the epilogue to *Mimesis*, he notes the difficulties of dealing with “texts ranging over three thousand years” and how the limitations of his library in Istanbul made it “probable that [he] overlooked things which [he] ought to have considered.” Interestingly, however, he says at the same time that “it is quite possible that the book owes its existence to just this lack of a rich . . . library.” If it had been possible to access the greater archive, he “might never have reached the point of writing” (1953).

fiction published in the eighteenth century? What of the works that Watt did not observe and account for with his methodology, and how are we to now account for the works not penned by Defoe, by Richardson, or by Fielding? Might other novelists tell a different story? Can we, in good conscience, even believe that Defoe, Richardson, and Fielding are representative writers? Watt's sampling was not random; it was quite the opposite. But perhaps we only need to believe that these three (male) authors are representative of the trend toward "realism" that flourished in the nineteenth century. Accepting this premise makes Watt's magnificent synthesis into no more than a self-fulfilling project, a project in which the books are stacked in advance. No matter what we think of the sample, we must question whether in fact realism really did flourish. Even before that, we really ought to define what it means "to flourish" in the first place. Flourishing certainly seems to be the sort of thing that could, and ought, to be measured. Watt had no such yardstick against which to make a measurement. He had only a few hundred texts that he had read. Today, things are different. The larger literary record can no longer be ignored: it is here, and much of it is now accessible.

At the time of my Thanksgiving dinner back in the 1990s, gathering literary evidence meant reading books, noting "things" (a phallic symbol here, a biblical reference there, a stylistic flourish, an allusion, and so on) and then interpreting: making sense and arguments out of those observations.* Today, in the age of digital libraries and large-scale book-digitization projects, the nature of the "evidence" available to us has changed, radically. Which is not to say that we should no longer read books looking for, or noting, random "things," but rather to emphasize that massive digital corpora offer us unprecedented access to the literary record and invite, even demand, a new type of evidence gathering and meaning making. The literary scholar of the twenty-first century can no longer be content with anecdotal evidence, with random "things" gathered from a few, even "representative," texts.† We must strive to understand these things we find interesting in the context of everything else, including a mass of possibly "uninteresting" texts.

* Yes, a simplification, but close enough to serve as a heady foil in this introductory polemic. Along similar lines, Susan Hockey writes of the "somewhat serendipitous noting of interesting features" (2000, 66).

† When writing of "anecdotal" here, I am not thinking of the use made of anecdote in the new historical tradition that we find expressed in, for example, Greenblatt's "cultural poetics." Rather, I use the word in the sense of "anecdotal evidence": that is, evidence that is atypical, informally gathered, speculative, or purely interpretive, which is to say not empirical. On this point, the type of literary data I am exploring allows me to adopt a fundamentally empirical position. Having said that, there is a place for anecdotal evidence in literary study, and I do not intend here a critique of anecdotalism per se, but rather to simply make a distinction and separation between two types of evidence.

“Strictly speaking,” wrote Russian formalist Juri Tynjanov in 1927, “one cannot study literary phenomena outside of their interrelationships” (1978, 71). Unfortunately for Tynjanov, the multitude of interrelationships far exceeded his ability to study them, especially with close and careful reading as his primary tools. Like it or not, today’s literary-historical scholar can no longer risk being *just* a close reader: the sheer quantity of available data makes the traditional practice of close reading untenable as an exhaustive or definitive method of evidence gathering. Something important will inevitably be missed. The same argument, however, may be leveled against the macroscale; from thirty thousand feet, something important will inevitably be missed. The two scales of analysis, therefore, should and need to coexist. For this to happen, the literary researcher must embrace new, and largely computational, ways of gathering evidence. Just as we would not expect an economist to generate sound theories about the economy by studying a few consumers or a few businesses, literary scholars cannot be content to read literary history from a canon of a few authors or even several hundred texts. Today’s student of literature must be adept at reading and gathering evidence from individual texts and equally adept at accessing and mining digital-text repositories. And *mining* here really is the key word in context. Literary scholars must learn to go beyond search. In search we go after a single nugget, carefully panning in the river of prose. At the risk of giving offense to the environmentalists, what is needed now is the literary equivalent of open-pit mining or hydraulicking. We are proficient at electronic search and comfortable searching digital collections for some piece of evidence to support an argument, but the sheer amount of data now available makes search ineffectual as a means of evidence gathering. Close reading, digital searching, will continue to reveal nuggets, while the deeper veins lie buried beneath the mass of gravel layered above. What are required are methods for aggregating and making sense out of both the nuggets and the tailings. Take the case of a scholar conducting research for a hypothetical paper about Melville’s metaphysics. A query for *whale* in the Google Books library produces 33,338 hits—way too broad. Narrowing the search by entering *whale* and *god* results in a more manageable 3,715 hits, including such promising titles as *American Literature in Context* and *Melville’s Quarrel with God*. Even if the scholar could further narrow the list to 1,000 books, this is still far too many to read in any practical way. Unless one knows what to look for—say, a quotation only partially remembered—searching for research purposes, as a means of evidence gathering, is not terribly practical.* More interesting, more exciting, than panning for nuggets in digital archives

* In revising this section before publication, I went back to Google Books and discovered that the number of hits for this particular search had grown significantly since I first tested. No doubt readers will find even higher numbers today.

is the ability to go beyond the pan and exploit the trommel of computation to process, condense, deform, and analyze the deeper strata from which these nuggets were born, to unearth, for the first time, what these corpora really contain. In practical terms, this means that we must evolve to embrace new approaches and new methodologies designed for accessing and leveraging the electronic texts that make up the twenty-first-century digital library.

This is a book about evidence gathering. It is a book about how new methods of analysis allow us to extract new forms of evidence from the digital library. Nevertheless, this is also a book about literature. What matter the methods, so long as the results of employing them lead us to a deeper knowledge of our subject? A methodology is important and useful if it opens new doorways of discovery, if it teaches us something new about literary history, about individual creativity, and about the seeming inevitability of influence.

3 TRADITION

Talents imitate, geniuses steal.

—[Oscar Wilde?]

As noted previously, there is a significant tradition of researchers employing computational approaches to the study of literature and an even longer tradition of scholars employing quantitative and statistical methods for the analysis of text. The specifically computational tradition dates back to the work of Father Roberto Busa, and since that time momentum has been building, exponentially, so that now, somewhat suddenly, the trend line has rocketed upward and the “digital humanities” have burst upon the scene and become a ubiquitous topic of discussion in humanities programs across the globe.* Notwithstanding the fact that there is no general agreement as to what exactly the term *digital humanities* defines, the sudden popularity of this thing called digital humanities has occurred with such rapidity that even we who consider ourselves natives of the tribe have been taken by surprise. Some have suggested that the reason stock in digital humanities is skyrocketing is because literary studies are in a general state of crisis and that we are yearning for a new theoretical construct that would ground our inquiries in science (see, for example, Gottschall 2008). This may be the case, for some, but I am not a member of that club. As someone who has studied diasporas, I understand that there can be pushes and pulls to any migration. For the Irish, British oppression made for an imposing stick and the promise of opportunity in America an enticing carrot. Here, however, the migration to digital humanities appears to be mostly about opportunity. In fact, the sudden motivation for scholars to engage in digital humanities is more than likely a direct by-product of having such a wealth of digital material with which to engage. With apologies to the indigenous, I must acknowledge here

* The tradition may stretch even further if we broaden our definition of *computation* to include substrates beyond silicon.

that the streets of this “new” world are paved with gold and the colonizers have arrived. A large part of this change in scholarly thinking about the digital has been brought about because of the very simple fact that digital objects, digital data stores, and digital libraries in particular have become both large and easily accessible. We have built it, and they are coming. Despite the success of this “thing called digital humanities,” as William Deresiewicz derided it in 2008, there remains no general agreement or even general understanding of what the term means or describes. Some, including Matthew Kirschenbaum (2010), think that this ambiguity is a good thing. I am not as certain. Do video-game analysis and stylometry really make good bedfellows? Probably not; these are entirely different threads.* Understanding how we got to this point of free-loving digital humanities is useful not simply as a matter of disciplinary history but as a way of contextualizing and understanding the methods and results presented in this book. So, a few words are in order about the traditions informing my macroanalytic approach to digital literary studies.

• • •

In 2012 we stand upon the shoulders of giants, and the view from the top is breathtaking. The skies were not always this clear. Susan Hockey summarized the period of the 1980s as one in which “we were still at a stage where academic respectability for computer-based work in the humanities was questionable” (2004, 10). Mark Olsen noted in 1993 that despite advances in text processing, “computerized textual research has not had a significant influence on research in the humanistic disciplines” (309). A decade later, Thomas Rommel argued that “the majority of literary critics still seem reluctant to embrace electronic media as a means of scholarly analysis . . . [and] literary computing has, right from the very beginning, never really made an impact on mainstream scholarship” (2004, 92). Stephen Ramsay wrote in 2007, “The digital revolution, for all its wonders, has not penetrated the core activity of literary studies, which, despite numerous revolutions of a more epistemological nature, remains mostly concerned with the interpretive analysis of written cultural artifacts. Texts are browsed, searched, and disseminated by all but the most hardened Luddites in literary study, but seldom are they transformed algorithmically as a means of gaining entry to the deliberately and self-consciously subjective act of critical interpretation” (478).

* Just so it is clear, I am a big fan of the “big tent,” or the “big umbrella,” if you will. In 2011 Glen Worthey and I cohosted the annual Digital Humanities Conference at Stanford, where our conference theme was “Big Tent Digital Humanities.” In the spirit of the Summer of Love, we donned tie-dyed shirts and let a thousand DH flowers bloom. We love our DH colleagues one and all. This book, however, stands at one side of the tent. We do different things in DH; we are vast.

Others from outside the scholarly community of computing humanists, writers such as Sven Birkerts (1994) and Nicholson Baker (2001), have warned of the dangers inherent in the digitization of books, and Emory English professor Mark Bauerlein has offered a sustained, if unspecific, critique of the digital age in general (2008). Even as recently as 2008, the ever-adversarial William Deresiewicz wrote in the *Nation* about the digital humanities, poking fun at something he imagined to be just another fad of scholarship.* But things change.

Despite some early concerns and several contemporary detractors, today—some few years after the most recent lamentations—the scholarly presses and the mainstream media are buzzing with news of this thing called “digital humanities.”† Humanities computing, or, more popularly, “digital humanities,” is alive and well. The field is healthy: participation in the primary professional organization, the Alliance of Digital Humanities Organizations (ADHO), is vibrant, and attendance at the annual Digital Humanities Conference is at an all-time high.‡ So large have we grown, in fact, that the number of rejected papers now far exceeds the number accepted, and many of the panels and papers that are not rejected draw standing-room crowds and lively discussion. Meanwhile, new degree programs specifically geared toward digital humanities are now offered at universities across the globe.§ Academic jobs for candidates with expertise in the intersection between the humanities and technology are becoming more and

* Wendall Piez offers an interesting response to Deresiewicz’s comment in “Something Called ‘Digital Humanities’” (2008).

† Matthew Kirschenbaum provides a succinct, six-page overview of the field in his *ADE Bulletin* article titled “What Is Digital Humanities and What’s It Doing in English Departments?” (2010). Other examples include Fischman 2008a, 2008b; Goodall 2008; Howard 2008a, 2008b, 2008c; Pannapacker 2011; Parry 2010; Shea 2008; and Young 2009.

‡ The Alliance of Digital Humanities Organizations is a consortium including the Association for Computing and the Humanities, the Association of Literary and Linguistic Computing, the Society for Digital Humanities, and CenterNet.

§ In terms of numbers of institutions per capita and dollars per capita, Canada is the obvious front runner here, but several universities in the UK, Ireland, and the United States have recently begun programs or “tracks” in digital humanities. Stanford began offering an undergraduate emphasis in “digital humanities” through its Interdisciplinary Studies in the Humanities Program back in 2006. In October 2006, Kings College of London announced a Ph.D. in digital humanities. In 2010 the National University of Ireland, Maynooth, began offering a master’s of arts in digital humanities (<http://www.learndigitalhumanities.ie/>), and University College London began offering a master of arts and science in digital humanities (<http://www.ucl.ac.uk/dh-blog/2010/07/30/announcing-the-new-mamsc-in-digital-humanities-at-ucl/>). In 2011 Trinity College Dublin began a master of philosophy program in digital humanities under the direction of Susan Schreibman.

more common, and a younger constituent of digital natives is quickly overtaking the aging elders of the tribe.* By one measure, the number of young scholars and graduate students attending the annual digital humanities conference in 2009 was three times the number of those attending one year earlier.† To my 2006 query to the members of the Humanist List about the health of the field, I received a number of encouraging replies that included remarks about the recent “groundswell of research interest” in digitally oriented projects, the development of new “centers” for computing in the humanities, and institutional support for the hiring of computing humanists.‡ Especially impressive has been the news from Canada. Almost all of the “G 10” (that is, the top thirteen research institutions of Canada) have institutionalized digital humanities activities in the form of degrees such as Alberta’s master’s in digital humanities, programs such as McMaster’s in digital media, centers such as the University of Victoria’s Humanities Computing Centre, or through institutes such as Victoria’s Digital Humanities Summer Institute. Noteworthy too is that the prestigious Canada Research Chair has been appointed to a number of computing humanists.§ Not the least important, the program for the 2011 Modern Language Association conference in Seattle included, by one scholar’s count, at least fifty-seven panels in the “digital humanities,” up from forty-four the previous year when the panel session titled “The History and Future of the Digital Humanities” had standing-room crowds (Pannapacker 2011).¶ All signs indicate that the digital

* A search, conducted in October 2006, of jobs listed in the *Chronicle of Higher Education* including both the words *digital* and *humanities* resulted in thirty-four hits. Recent searches have contained even more, including opportunities in senior-level posts such as that advertised in July 2010 for a director of Texas A&M’s new Digital Humanities Institute. On September 25, 2011, Desmond Schmidt posted the following summary of digital humanities jobs on the Humanist List: “There have been a lot of advertisements for jobs lately on Humanist. So I used the Humanist archive to do a survey of the last 10 years. I counted jobs that had both a digital and a humanities component, were full time, lasted at least 12 months and were at PostDoc level or higher. 2002: 11, 2003: 6, 2004: 15, 2005: 15, 2006: 18, 2007: 24, 2008: 27 (incomplete - 1/2 year), 2009: 36, 2010: 58. 2011: 65 so far.”

† In 2009 I was chair of the ADHO Bursary Awards committee. The prize is designed to encourage new scholars in the discipline. From 2008 to 2009, the number of candidates for the Bursary Award jumped from seven to more than thirty.

‡ Humanist, now in its twenty-second year of operation, is, by general consensus, the Listserv of record for all matters related to computing in the humanities.

§ See http://tapor.ualberta.ca/taporwiki/index.php/Canada_Research_Chairs_and_Award_Winners.

¶ See Mark Sample, <http://www.samplereality.com/2011/10/04/digital-humanities-sessions-at-the-2012-mla-conference-in-seattle/> and <http://www.samplereality.com/2010/11/09/digital-humanities-sessions-at-the-2011-mla/>.

humanities have arrived, even while the fields of study sheltering beneath the umbrella remain a somewhat ambiguous and amorphous amalgamation of literary formalists, new media theorists, tool builders, coders, and linguists.

Computational text analysis—by all accounts the foundation of digital humanities and its deepest root—has come a long way since 1949, when Father Roberto Busa began creation of his word index. These days, humanists routinely create word indexes and frequency lists using readily available software. With the spread of broadband and the accessibility of the Internet, many tools that were once platform dependent and command line in nature have been “reinvented” for the web so that scholars may now do small-scale text processing and analysis on remote web servers using any number of web-based applications. Keyword-in-context lists can be quickly generated using TactWeb.* Stéfan Sinclair’s HyperPo and Voyant offer self-serve text-analysis tools for traditional concordancing and co-occurrence alongside more experimental widgets for the processing and deforming of textual data.† There is a growing number of tools specifically geared toward the “visualization” of literary materials.‡ A particularly well-conceived, low-entry project is the “Text Analysis Portal” (TAPoR), which has set itself up as a one-stop shop for basic text analysis. This project, which began life with a six-million-dollar (CAD) grant from the Canadian Foundation for Innovation, is distributed across six universities and provides a centralized and, to some extent, standardized way of accessing a variety of text-analysis applications. TAPoR serves as a model of collaboration and offers a foundational, even seminal, approach to future humanities computing work. Indeed, some in the United States are now attempting to go beyond TAPoR and develop what Chris Mackey, formerly of the Mellon Foundation, once referred to as the “mother of all text-analysis applications.”§ These projects, whose names include “Bamboo” and others with such funky acronyms as MONK, SEASR, and DARIAH, are all seeking ways to make leveraging computation as easy for the average literary scholar as finding biblical references in a canonical novel.¶

* See <http://tactweb.humanities.mcmaster.ca/tactweb/doc/tact.htm>.

† See <http://tapor1.mcmaster.ca/~sgs/HyperPo/> and <http://voyant-tools.org/>.

‡ See, for example, Bradford Paley’s TextArc application (<http://www.textarc.org/>) or the word clouds available through Wordle or the Many Eyes project of IBM.

§ The comment was made during a presentation at the Stanford Humanities Center. The project that eventually emerged from these and other discussions is Project Bamboo: <http://www.projectbamboo.org>.

¶ MONK stands for “Metadata Offers New Knowledge” (<http://www.monkproject.org/>), SEASR for Software Environment for the Advancement of Scholarly Research (<http://seasr.org/>), and DARIAH for Digital Research Infrastructure for the Arts and Humanities (<http://www.dariah.eu/>). See also Project Bamboo at <http://www.projectbamboo.org>.

Computing humanists have made important contributions to humanities scholarship: thanks to them, we have impressive digital archives and critical editions such as the exemplary Women Writers Project of Brown University and Kevin Kerinan's impressive Electronic Beowulf.* Fellow travelers from linguistics, machine learning, natural language processing, and computer science have developed robust text-analysis programs that can be employed to automatically identify parts of speech, named entities (people, places, and organizations), prominent themes, sentiment, and even poetic meter.† These tools have in turn been deployed for studies in authorship attribution, textual dating, and stylistic analysis.

There are any number of other useful products that have evolved out of collaborations among humanists, linguists, and technologists: the Google search engine performs a type of text analysis when searching for keywords and collocates; using calculations based on vocabulary, sentence length, and syllables, Microsoft Word attempts to determine the grade level of a piece of writing.‡ The XML (extensible markup language) standard that plays such a critical role in data interchange today was heavily influenced by the early work of the Text Encoding Initiative and in particular founding TEI editor Michael Sperberg-McQueen. These have been important and useful contributions, to be sure, and the recent Blackwell publications *A Companion to Digital Humanities* (Schreibman, Siemens, and Unsworth 2004) and *A Companion to Digital Literary Studies* (Siemens and Schreibman 2007) are a testament to the various ways in which technology has established itself in the humanities.

Despite all of this achievement and the overwhelming sense of enthusiasm and collegiality that permeates the DH community, there is much more work to be done. We have in fact only begun to scratch the surface of what is possible. Though the term *digital humanities* has become as omnipresent on our campuses as *multiculturalism* was several years ago, the adoption of “digital” tools and methodologies has been limited, even among those who would self-identify as “digital humanists.” To be sure, literary scholars have taken advantage of digitized textual material, but this use has been primarily in the arena of search, retrieval, and access. We have not yet seen the scaling of our scholarly questions in accordance with the massive scaling of digital content that is now

* <http://www.wwp.brown.edu/> and <http://ebeowulf.uky.edu/>.

† Examples include the Stanford Natural Language Processing Group's Part of Speech Tagger and Named Entity Recognizer, the University of Massachusetts's Machine Learning for Language Toolkit (MALLET), and many others.

‡ For more on this, just open Microsoft Word's “Help” and search for “Readability Scores.” MS Word uses both the Flesch Reading Ease score and the Flesch-Kincaid Grade Level score.

held in twenty-first-century digital libraries. In this Google Books era, we can take for granted that some digital version of the text we need will be available somewhere online, but we have not yet fully articulated or explored the ways in which these massive corpora offer new avenues for research and new ways of thinking about our literary subject.*

To some extent, our thus-far limited use of digital content is a result of a disciplinary habit of thinking small: the traditionally minded scholar recognizes value in digital texts because they are individually searchable, but this same scholar, as a result of a traditional training, often fails to recognize the potentials for analysis that an electronic processing of texts enables. For others, the limitation is more directly technical and relates to the type and availability of software tools that might be deployed in analysis. The range of what existing computer-based tools have provided for the literary scholar is limited, and these tools have tended to conform to a disciplinary habit of closely studying individual texts: that is, close reading. Such tools are designed with the analysis of single texts in mind and do not offer the typical literary scholar much beyond advanced searching capabilities. Arguably, the existing tools have been a determiner in shaping perceptions about what can and cannot be done with digital texts.† The existing tools have kept our focus firmly on the close reading of individual texts and have undoubtedly prevented some scholars from wandering into the realms of what Franco Moretti has termed “distant reading” (2000). Combine a traditional literary training focused on close reading with the most common text-analysis tools focused on the same thing, and what you end up with is enhanced search—electronic finding aids that replicate and expedite human effort but bring little to the table in terms of new knowledge. I do not intend to demean the use of text-analysis tools at the scale of the single text or at the scale of several texts; quite the contrary, there is an incredibly large body of quantitative work in authorship attribution, gender identification, and what is

* My comments here may seem idealistic given the realities of copyright law and contemporary literature in particular. That digital versions of these recent works exist seems a point we can take for granted; that they are or will be readily accessible is a more complicated problem about which I have more to say in chapter 10.

† Duke University historian of science Tim Lenoir has made a similar point in arguing that quarks would not exist were it not for the particle accelerators that were built to discover or produce them. Lenoir has made this comment on multiple occasions, primarily in lectures on pragmatic realism and social construction. He has written about this extensively in his book *Instituting Science* (1997), particularly the chapter on Haber-Bosch, in which he discusses this issue at length. He derived this line of thinking in part from Ian Hacking’s argument in *Representing and Intervening*, in which Hacking argues that electrons are real when you can spray them (1983, 23).

more generally referred to as “stylometry” that informs my own work. And even in the less statistically driven realms of computational text analysis, there are tools for visualizing and exploring individual texts that serve as rich platforms for “play,” as Stéfan Sinclair has termed it (2003), or what might more formally be termed “discovery” and “exploration.” Steven Ramsay’s “Algorithmic Criticism” (2007) provides a strong statement regarding the value of text-analysis tools for text “deformation.” Such deformations may lead to new and different interpretations and interpretive strategies.*

Our colleagues in linguistics have long understood the value of working with large corpora and have compiled such valuable resources as the British National Corpus and the Standard Corpus of Everyday English Usage. Linguists employ these resources in order to better understand how language is used, is changing, is evolving. The tools employed for this work are not, generally speaking, web-based widgets or text-analysis portals such as the TAPoR project. Instead, our colleagues in linguistics have learned to be comfortable on the command line using programming languages. They have learned to develop applications that run on servers, and they have developed a willingness to wait for their results. Literary scholars, on the other hand, have generally been content to rely upon the web for access to digital material. Even in the text-analysis community, there is a decided bias in favor of developing web-based tools.† Unfortunately, the web is not yet a great platform upon which to build or deliver tools for doing text analysis “at scale.” Quick queries of indexed content, yes, but not corpus ingestion or complex analysis.‡

Given the training literary scholars receive, their typical skill set, and the challenges associated with large-scale digitalization and computational analysis, it is easy to understand why literary scholars have not asked and probed with computers the same sorts of questions about “literary language” that linguists

* Ramsay’s original article has now been extended into a book-length study. See Ramsay 2011.

† Stéfan Sinclair of McGill University is an accomplished text-analysis tool builder, and his recent offering, Voyant, is the best example I have seen of an online tool that can handle a large amount of text. See <http://voyant-tools.org/>. Even this exceptional tool is still only capable of fairly basic levels of analysis.

‡ Cloud computing and high-performance computing are certainly beginning to change things, and projects such as SEASR may someday provide the web interface to high-performance text analysis. At least in the near term, the success of web-based macroanalysis will depend in large part upon the users of such tools. They will need to abandon the idea that clicking a link returns an immediate result. The web may become a portal into a complex text-analysis platform, but the web is not likely to evolve as a place for instant access to complex data.

have asked about language in general. On the one hand, literary scholars have not had access, until recently, to large amounts of digital literary content, and, on the other, there is a long-standing disciplinary habit of thinking about literature in a limited way: in terms of “close readings.” Close reading is a methodological approach that can be applied to individual texts or even small subsets of texts but not, for example, to all British fiction of the nineteenth century. A “close reading” of nineteenth-century British fiction would, in fact, be implausible. Consider, for example, the very real limitations of human reading: Franco Moretti has estimated that of the twenty to thirty thousand English novels published in Britain in the nineteenth century, approximately six thousand are now extant. Assuming that a dedicated scholar could find these novels and read one per day, it would take sixteen and a half years of close reading to get through them all. As a rule, literary scholars are great synthesizers of information, but synthesis here is inconceivable.* A computer-based analysis or synthesis of these same materials is not so difficult to imagine. Though the computer cannot perfectly replicate human synthesis and intuition, it can take us a long way down this road and certainly quite a bit further along than what the human mind can process. It is exactly this kind of macroanalytic approach that is the future of computing in the humanities, and, according to some, the future of literary studies (see, for example, Gottschall 2008 and Martindale 1990).

I am not the first, however, to suggest that a bird’s-eye view of literature might prove fruitful. On this point, Franco Moretti has been at the forefront, suggesting “distant reading” as an alternative to “close reading.” In *Graphs, Maps, Trees*, Moretti writes of how a study of national bibliographies made him realize “what a minimal fraction of the literary field we all work on: a canon of two hundred novels, for instance, sounds very large for nineteenth-century Britain (and is much larger than the current one), but is still less than one per cent of the novels that were actually published: twenty thousand, thirty, more, no one really knows—and close reading won’t help here, a novel a day every day of the year would take a century or so” (2005, 3–4). Moretti’s “Graphs” chapter is particularly compelling; it provides a beginning point for the development

* In history and in historical economics, there is a recent tradition of thinking big. The *Annales* school of historiography developed by the French in the early twentieth century has had the goal of applying quantitative and social-scientific methods in order to study history of the “long-term,” the *longue durée*. The approach views history in terms of “systems.” Lynn Hunt’s brief and useful overview of the history of the *Annales* paradigm argues that “in contrast to earlier forms of historical analysis [namely, exemplar and developmental approaches], the *Annales* school emphasized serial, functional, and structural approaches to understanding society as a total, inter-related organism” (1986, 211).

of a more formal literary time-series analysis methodology. Moretti examines the publication rates for novels (in several countries) over periods of years and decades. Focusing on the peaks and valleys in novel production, he moves from the quantitative facts to speculation and interpretation, posing, for example, that the rise and fall of various novelistic genres in the British corpus can be correlated to twenty-five- to thirty-year cycles or generations of readers. In Moretti's model, the tastes and preferences of one generation are inevitably replaced by those of the next. He suggests that there are connections between literary cycles and political ones, arguing, for example, that the French Revolution was a critical factor in the fall of the French novel. Although such an argument could certainly be made anecdotally, the accompanying data—and the graph showing the sharp decline in novel production in about 1798—leave little room for debate.

Nor am I original in considering the applications of technology to large textual collections. Already noted are the linguists, and there is, of course, an entire community of computer scientists (many of them at Google) who work in the field of text mining and information retrieval. Along with similar agencies in other nations, the National Security Agency is in this business as well: the NSA is reported to have been employing text-mining technologies since the Cold War, and the “classified” ECHELON surveillance system is purported to capture all manner of electronic information, from satellite communications to email correspondences. These captured materials are then analyzed, mined by machines, in order to sniff out threats to national security. The amount of information devoted to ECHELON online is somewhat staggering—a Google search for this supersecret program along with the keywords *text* and *mining* provides 375,000 sites of interest. This figure is trivial next to the Google results for a search for the keyword *Area 51* (154 million hits) but does demonstrate the point that text mining, and ECHELON for that matter, is nothing new. Similar to ECHELON is the technology developed by Palantir Technologies in Palo Alto, California. The company's website describes their software as being “a platform for information analysis . . . designed for environments where the fragments of data that . . . tell a larger story are spread across a vast set of starting material” (Palantir Technologies 2011). Translation: we build technologies for the macroanalysis of large, disparate corpora.

Not quite as spectacular as Palantir and the NSA are projects more specifically aimed at the application of text mining to the humanities. The NORA, MONK, and SEASR projects originally led by John Unsworth at the University of Illinois are three such projects. The expressed goal of the NORA project was to “produce software for discovering, visualizing, and exploring significant patterns across large collections of full-text humanities resources in existing digital libraries” (NORA 2006). Using software developed by the University of Illinois's

National Center for Supercomputing Applications and the “Data to Knowledge” applications of Michael Welge’s Automated Learning Group, the NORA team successfully deployed a Java-based application for “sniffing” out preidentified “patterns” in large digital collections. An early version of the software allowed an end user to “tag” or “mark” certain works in a collection, and the system then used those works to build a model—what some biologists who work with DNA and gene expression call a “signal.” This signal is then sought throughout the larger collection. The example offered on the NORA website involves marking “erotic” passages in the works of Emily Dickinson. Some 260 individual documents are presented, and the user “marks” or rates a small percentage of these for erotic content.* The human-marked documents constitute a training set, which is used by the software to “predict” which works in the collection are likely to contain erotic content as well. This is essentially an information-retrieval task. MONK and SEASR are more advanced implementations of the NORA technologies. SEASR provides the most deeply abstracted and robust imagining of the early NORA work. SEASR is both a back-end infrastructure and a semifriendly web interface that allows researchers to build text-analysis “flows” that get executed on a server.†

Outside of the humanities, computer scientists working in natural language processing, corpus linguistics, and computational linguistics have developed a wide range of tools that have direct application to work in literary studies. Using a technique called “topic modeling,” a group led by David Newman at the University of California–Irvine (UCI) harvested the latent themes, or topics, contained in 330,000 stories published in the *New York Times*. The topic-modeling procedure they employed required no human preprocessing; it was “unsupervised” in its sifting through a corpus of documents and then identifying patterns of words that were frequently collocated.‡ The software categorizes the words in each document into mathematically correlated clusters, which are described as “topics.” Not surprisingly, the UCI team first presented their research at the Intelligence and Security Informatics conference in San Diego (Newman, Smyth, and Steyvers 2006). More interesting (for scholars of literature) than the

* This process of human intervention is known in data and text mining as “supervised learning.”

† From 2011 to 2012, I served as the project lead on “Phase Two” of the SEASR project. The work was generously funded by the Mellon Foundation.

‡ Andrew McCallum and his team at the University of Massachusetts have done exciting work developing a “Machine Learning for Language Toolkit,” or “MALLET,” which provides functionality for a variety of text-mining applications. The MALLET software includes David Mimno’s topic-modeling code, which is used and described at length in chapter 8.

intelligence applications of topic modeling are the applications to humanities research. Historian Sharon Block, for example, teamed up with Newman and employed topic-modeling routines to explore the entire eighteenth-century run of the *Pennsylvania Gazette*. In her essay “Doing More with Digitization: An Introduction to Topic Modeling of Early American Sources” (2006), Block walks readers through a series of examples of how the technique can assist historians and reveal new avenues for research in the form of unanticipated patterns and trends.* Though not designed with literary scholarship in mind, the topic-modeling tools can be applied to literary texts at the level of the corpus or even at the level of the individual book or poem.†

Still another project working to apply the tools and techniques of text mining and corpus linguistics to literature is the WordHoard project at Northwestern University. Ironically, the WordHoard site describes its software as “an application for the close reading and scholarly analysis of deeply tagged texts” but then goes on to say that it “applies to highly canonical literary texts the insights and techniques of corpus linguistics, that is to say, the empirical and computer-assisted study of large bodies of written texts or transcribed speech” (WordHoard 2006). The descriptive prose that follows adds that the software allows for a deeply “microscopic” and philological inquiry of the text(s). Although it is true that WordHoard provides access to, or tools for, harvesting richly encoded texts, the results being gleaned from the texts are not so much the results of a close reading–like process as they are the results of a macroscopic text-mining process that aggregates a number of relatively small details into a more global perspective. As such, the process seems to have less in common with close-reading practices and more with Moretti’s notion of distant reading. The devil is in the details and in how the details are investigated and aggregated in order to enable a larger perspective. Writing of “detailism” and digital texts, Julia Flanders discusses Randolph Starn’s introduction to a special issue of

* Cameron Blevins provides another historical example. Blevins uses topic modeling to explore entries in Martha Ballard’s eighteenth-century diary. See <http://history.org/2010/04/01/topic-modeling-martha-ballards-diary/>.

† David Newman was the first guest speaker in the Beyond Search workshop that I ran at Stanford from 2006 to 2009. Prior to his arrival, I prepared a corpus of texts for Newman to process. Included in those data were the novels of Jane Austen. As part of his presentation, Newman showed how the topic of “sentiment” (composed of words denoting emotion) could be tracked throughout the Austen corpus. Looking at the graphs that he prepared, participants in the workshop could see how Austen employs moments of strong emotion throughout her texts. In some novels, we observed a regular fluctuation, while others showed a steady trend upward: as the novels progressed, the presence of strong emotions increased.

Representations. She notes the effort to connect “detail . . . with a larger historical view.” She goes on to emphasize that detail is used “not as ‘mere facts’ cited as evidence . . . but as the contextually embedded ‘trace, clue, sign, shard’ that carries a specifiable, signifying linkage to some historical genealogy” to some larger system (2005, 43). WordHoard offers a way of aggregating these signs into a coherent argument. The website offers the word *love*—as it appears in the works of Chaucer, Spenser, and Shakespeare—as an example. Female characters, the data reveal, are “about 50% more likely to speak of love than men.” This conclusion is derived not through a computer-based close reading of the texts, but rather via a quantitative zooming out and away from the texts, a zooming out that allows the user to simultaneously “see” all of the separate occurrences of the word throughout the corpus. The end result is that the WordHoard tool takes us quite far away from the actual occurrences of the words in the texts; our attention is drawn to an examination of the bigger picture, the macroview of *love* when used as a noun, of *love* when used as a verb, and in both cases of *love* as it is used by male or female speakers. This is not close reading; this is macroanalysis, and the strength of the approach is that it allows for both zooming in and zooming out.*

* A relatively recent entry into this realm of micro-macro-oriented text-analysis tools is Aditi Muralidharan’s WordSeer (<http://wordseer.berkeley.edu>). Australian digital humanist Tim Sherratt offers another variety of similar tools via his “WraggeLabs Emporium” (<http://wraggelabs.com/emporium>).