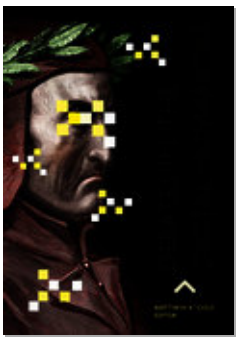
Debates in the Digital Humanities

Matthew K. Gold



Published by University of Minnesota Press

Canons, Close Reading, and the Evolution of Method

MATTHEW WILKENS

I have a point from which to start: canons exist, and we should do something about them. The digital humanities offer a potential solution to this problem, but only if we are willing to reconsider our priorities for digital work in ways that emphasize quantitative methods and the large corpora on which they depend.

I wouldn't have thought the first proposition, concerning canons and the need to work around them, was a dicey claim until I was scolded recently by a senior colleague who told me that I was thirty years out of date for making it. The idea being that we'd had this fight a generation ago, and the canon had lost. But I was right and he, I'm sorry to say, was wrong. Ask any grad student reading for her comps or English professor who might confess to having skipped Hamlet. As I say, canons exist. Not, perhaps, in the Arnoldian-Bloomian sense of *the* canon, a single list of great books, and in any case certainly not the *same* list of dead white male authors that once defined the field. But in the more pluralist sense of books one really needs to have read to take part in the discipline? And of books many of us teach in common to our own students? Certainly. These are canons. They exist.

So why, a few decades after the question of canonicity as such was in any way current, do we still have these things? If we all agree that canons are bad, why haven't we done away with them? Why do we merely tinker around the edges, adding a Morrison here and subtracting a Dryden there? What are we going to do about this problem? And more to the immediate point, what does any of this have to do with digital humanities and with debates internal to digital work?

The Problem of Abundance

The answer to the question "Why do we still have canons?" is as simple to articulate as it is apparently difficult to solve. We don't read any faster than we ever did, even as the quantity of text produced grows larger by the year. If we need to read books in order to extract information from them and if we need to have read things in common in order to talk about them, we're going to spend most of our time dealing

with a relatively small set of texts. The composition of that set will change over time, but it will never get any bigger. This is a canon.[1]

To put things in perspective, consider the scale of literary production over the last few decades as shown in Figure 14.1. Two things stand out: First, there are a lot of new books being published every year, and the number has grown rapidly over the last decade. Even excluding electronic editions and print-on-demand titles (as these figures do), we're seeing fifty thousand or more new works of long-form fiction annually in the United States alone (and at least as many again in the rest of the world at a time when national divisions are growing less relevant to cultural production). The overall U.S. market for books isn't growing beyond the larger economy (publishing revenues as a share of GDP have been constant, at about 0.2 percent, for decades [see Greco et al.]), but it's now being split among far more titles. This is likely the result of decreasing publishing costs in everything from acquisitions to distribution and marketing. The surge in quantity of published texts is surely a good thing insofar as it represents—in raw terms, at least—greater access to the market for a wider range of authors and a more diverse choice of books for readers. But it also means that each of us reads only a truly minuscule fraction of contemporary fiction (on the order of 0.1 percent, often much less). We could call this situation the problem of abundance. It is plainly getting worse with time.

We should notice, too—this is the second observation concerning Figure 14.1—that although the number of titles published annually was much lower prior to 2000,

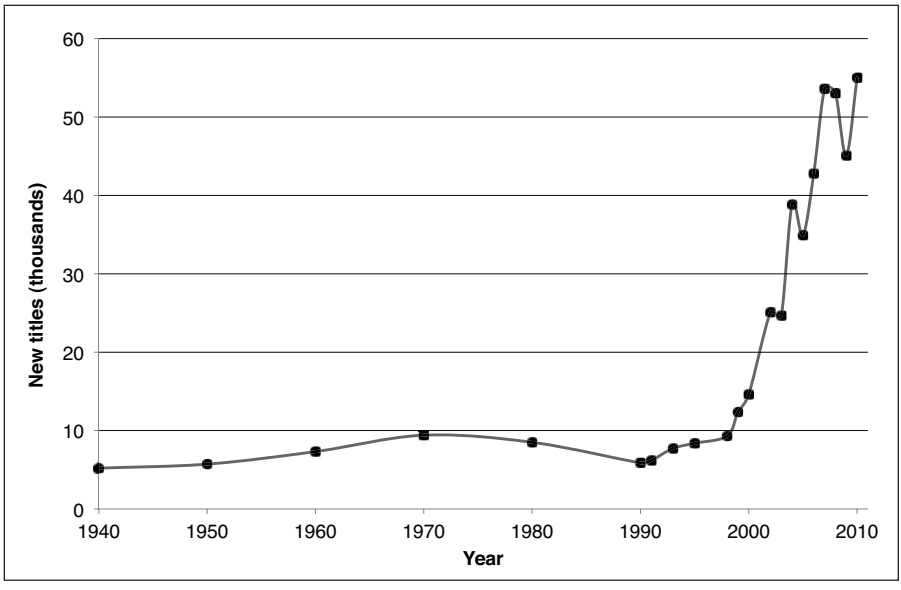


Figure 14.1. Number of new fiction titles published annually in the United States between 1940 and 2010. Sources: Greco et al. (1940–1991); R. R. Bowker (1993–2010). Title output for 1991 and earlier adjusted (upward, by a factor of three) to account for Bowker's subsequent changes in methodology.

it wasn't low in any absolute sense. There have been thousands of new novels published every year for decades. Looking further back, British publishers were bringing out more than a novel a week, on average, as early as the late eighteenth century (Moretti, 7, citing several studies); and American publishers managed the same feat by the mid-nineteenth century (Wright). So although the scale of the inadequacy of our reading is greater now than it was even a decade ago, the fact that we haven't been able to keep up with contemporary literary production is nothing new. We've had a version of the problem of abundance for centuries, whether we've cared to acknowledge it or not.

Another way of putting the issue would be to say that we need to decide what to ignore. And the answer with which we've contented ourselves for generations is, "Pretty much everything ever written." Even when we read diligently, we don't read very much. What little we do read is deeply nonrepresentative of the full field of literary and cultural production, as critics of our existing canons have rightly observed for decades. In the contemporary case, we read largely those few books that are actively promoted by the major publishing houses and in any case almost exclusively books that have been vetted through commercial publication. When we try to do better, to be more inclusive or representative in both form and content, we do so both with a deep set of ingrained cultural biases that are largely invisible to us and (because we've read so little) in ignorance of their alternatives. This is to say that we don't really have any firm sense of how we would even try to set about fixing our canons, because we have no idea how they're currently skewed. We're doing little better, frankly, than we were with the dead white male bunch fifty or a hundred years ago, and we're just as smug in our false sense of intellectual scope. The problem, moreover, is getting worse as the store of unread books grows with each passing week.

So canons—even in their current, mildly multiculturalist form—are an enormous problem, one that follows from our single working method as literary scholars—that is, from the need to perform always and only close reading as a means of cultural analysis. It's probably clear where I'm going with this, at least to some of those who work in digital literary studies. We need to do less close reading and more of anything and everything else that might help us extract information from and about texts as indicators of larger cultural issues. That includes bibliometrics and book historical work, data mining and quantitative text analysis, economic study of the book trade and of other cultural industries, geospatial analysis, and so on. Franco Moretti's work is an obvious model here, as is that of people like Michael Witmore on early modern drama and Nicholas Dames on social structures in nineteenth-century fiction.

Example: Text Extraction and Mapping

One example of what I have in mind is shown in Figure 14.2, which presents a map of the locations mentioned in a corpus of thirty-seven American literary texts

published in 1851. There are some squarely canonical works included in this collection, including *Moby Dick* and *House of the Seven Gables*, but the large majority are obscure novels by the likes of T. S. Arthur and Sylvanus Cobb. I certainly haven't read many of them, nor am I likely to spend months doing so. The texts are drawn from the Wright American Fiction Collection[2] and represent about a third of the total American literary works published that year.[3] The Wright collection is based on Lyle Wright's *American Fiction, 1851–1875: A Contribution Toward a Bibliography*, which attempts to identify and list every work of long-form fiction by an American author published in the United States between the relevant dates. Place names were extracted using a tool called Geodict, which looks for strings of text that match a large database of named locations.[4] A bit of cleanup on the extracted places was necessary, mostly because many personal names and common adjectives are also the names of cities somewhere in the world. I erred on the conservative side, excluding any such named places found and requiring a leading preposition for cities and regions. If anything, some valid places have likely been excluded. But the results are fascinating.

Two points of particular interest concerning the figure in this map: First, there are more international locations than one might have expected. True, many of them are in Britain and western Europe, but these are American novels, not British or other reprints, so even that fact might surprise us. And there are also multiple mentions of locations in South America, Africa, India, China, Russia, Australia, the Middle East, and so on. The imaginative landscape of American fiction in the mid-nineteenth century appears to be pretty diversely outward looking in a way that hasn't yet received much attention. Indeed, one of the defining features of our standard model of the period is that its fiction is strongly introspective at

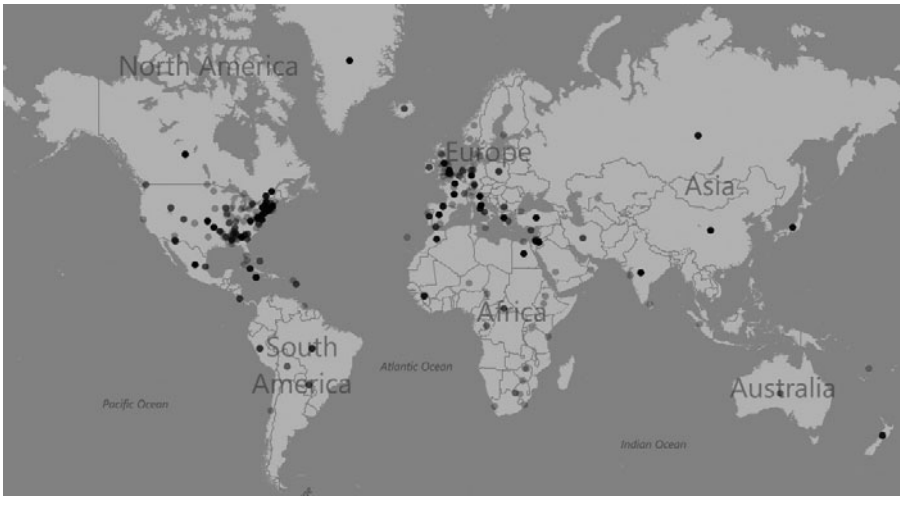


Figure 14.2. Places named in thirty-seven U.S. novels published in 1851.

both the personal and national levels, concerned largely with American identity and belonging.

Second, there's the distinct cluster of named places in the American South. At some level, this probably shouldn't be surprising, since we're talking about books that appeared just a decade before the Civil War, and the South was certainly on people's minds. But it doesn't fit very well with the stories we currently tell about Romanticism and the American Renaissance, which are centered firmly in New England during the early 1850s and dominate our understanding of the period. Perhaps we need to at least consider the possibility that American regionalism took hold significantly earlier than we usually claim.

How do these results compare with other years represented in the Wright corpus? Consider Figures 14.3 and 14.4. The maps for these years—1852 and 1874, respectively—are a bit noisier than 1851 because they've undergone less intensive data curation, but the patterns of location distribution are broadly similar to those observed earlier.[5] Two features stand out, however, in a comparison of locations appearing before and after the Civil War:

1. The density of named locations in the American west is noticeably greater in 1874 than in 1852.
2. The emergence of a second, distinct cluster of locations in the south-central United States, vaguely discernible in the earlier maps, is more pronounced in the 1874 map.

Both of these developments offer potential evidence for a revised understanding of American regionalism, particularly insofar as the dates at which the west and south-central regions become part of the imaginative landscape of American fiction significantly precede the presence of meaningful publishing activity in either area. We will want to know more about both the social history of westward expansion and the specific uses to which these locations are put in the period's literature before drawing any strong conclusions concerning their relevance, but it's clear that information of this kind can make up an important part of any new story about the course of American fiction in the nineteenth century.

There's also something to be said, however, about the striking overall similarity of all three maps. The Civil War is *the* periodizing event of American literary history. The books in question (more than a hundred from each period 1851–53 and 1873–75) were written roughly a generation apart and a decade before and after the war. If ever we should expect significant literary change in response to rapid social transformation, the Wright corpus is a strong candidate in which to observe it. So the question is, do we see a meaningful shift in a relevant aspect of literary production before and after the Civil War evinced in the data? To which the answer is, it depends on what we mean by meaningful. Our working image of important shifts in literary and cultural production is one derived from our experience with a

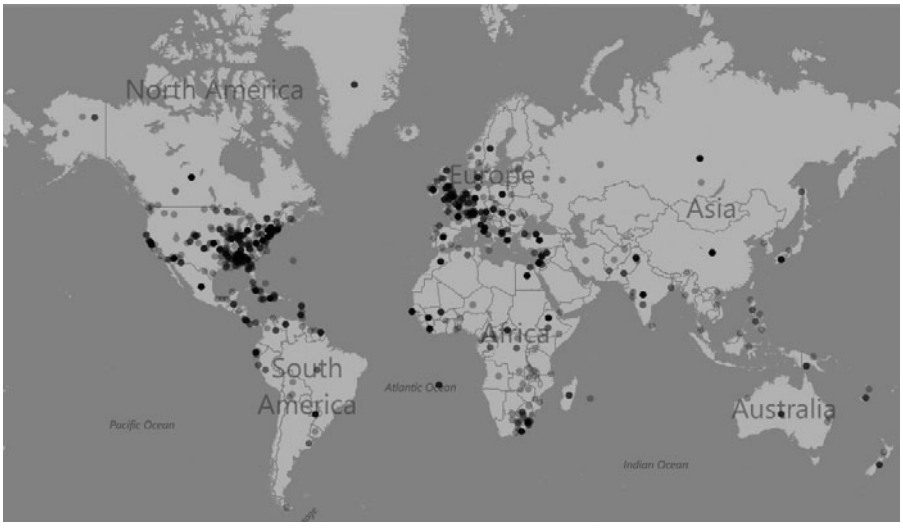


Figure 14.3. Places named in forty-five U.S. novels published in 1852.

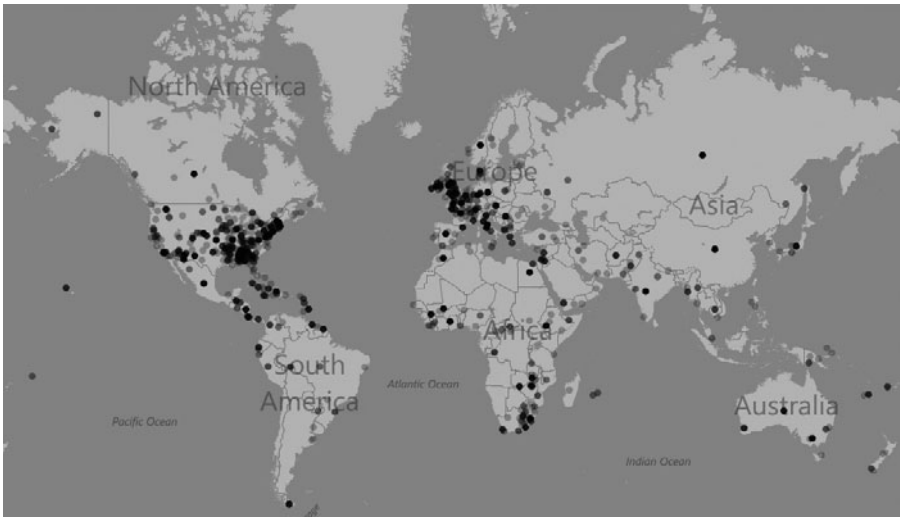


Figure 14.4. Places named in thirty-eight U.S. novels published in 1874.

handful of purportedly significant or representative works. We have, by working of necessity with very few texts and constructing highly detailed accounts of their particular differences, perhaps allowed ourselves to see period differences (and other classificatory distinctions) as more far reaching than they really are. This isn't to say that differences between periods, genres, nations, genders, and so on don't exist but only that they may consist in small but important variations on larger underlying continuities.

A scheme of literary classification based on small changes in overall literary production certainly isn't a repudiation of, for instance, periodization as such, but it does have serious implications concerning our work with the larger corpora that have long been outside our methodological grasp. As we begin to ask questions concerning the patterns and differences between and among hundreds or thousands of books, we will almost certainly encounter many more instances of data that resemble the maps presented here, in which we observe small variations in features that are broadly shared across corpora. One available interpretation of this fact—an interpretation based on our established model of event-based change—will be that no significant underlying change has occurred. Perhaps we will go so far as to say that literary history is best understood as an essentially unbroken chain of (at most) gradual evolutionary steps. Perhaps. But we should also be aware that we currently have very little concept of what major variations in the form or content of large bodies of literature might look like. It's certainly possible for small absolute variations of a given metric to be culturally (or physically, for that matter) significant in other contexts. Think, for instance, of the difference between a close national election and a comfortably one-sided vote. If our response to variations of, say, *only* 10 or 20 percent is that they fall far short of our idea of transformational, it may be our concept of transformation that needs to change.

One further methodological note: The techniques involved here can be readily extended with little additional time or labor to new texts as they become available. This is important because it allows us to test our hypothesis against new material and in different configurations and eras, something that would be difficult or impossible to do by hand. There's no guarantee, of course, that we'll be able to explain what we find or that it will fit neatly with what we've seen so far, but that's a feature, not a bug. It's good to have more material for evaluation and analysis. The relative speed of these methods is also an advantage insofar as it allows us to look for potentially interesting features without committing months or years to extracting them via close reading. We may very well still need to read some of the texts closely, but text-mining methods allow us to direct our scarce attention to those materials in which we already have reason to believe we will find relevant information. Though we're not used to framing our work in terms of rapid hypothesis testing and feature extraction, the process isn't radically different from what we already do on a much smaller scale. Speed and scalability are major benefits of this strand of computational work.

Consequences and Conclusions

I think the maps presented earlier offer interesting preliminary results, ones that demonstrate the first steps in a type of analysis that remains literary and cultural but that doesn't depend on close reading alone nor suffer the material limits such reading imposes. I think we should do more of this—not necessarily more geolocation

extraction in mid-nineteenth-century American fiction (though what I've shown obviously doesn't exhaust that project) but certainly more algorithmic and quantitative analysis of piles of text much too large to tackle "directly." ("Directly" gets scare quotes because it's a deeply misleading synonym for close reading in this context.)

If we do that—shift more of our critical capacity to such projects—there will be a couple of important consequences. For one thing, we'll almost certainly become worse close readers. Our time is finite; the less of it we devote to an activity, the less we'll develop our skill in that area. Exactly how much our reading suffers and how much we should care are matters of reasonable debate; they depend on both the extent of the shift and the shape of the skill-experience curve for close reading. My sense is that we'll come out all right and that it's a trade—a few more numbers in return for a bit less text—well worth making. We gain a lot by having available to us the kinds of evidence text mining (for example) provides, enough that the outcome will almost certainly be a net positive for the field. But I'm willing to admit that the proof will be in the practice and that the practice is, while promising, as yet pretty limited. The important point, though, is that the decay of close reading as such is a negative in itself only if we mistakenly equate literary and cultural analysis with their current working method.

Second—and here's the heart of the debate for those of us already engaged in digital projects of one sort or another—we'll need to see a related reallocation of resources within DH itself. Over the last couple of decades, many of our most visible projects have been organized around canonical texts, authors, and cultural artifacts. They have been motivated by a desire to understand those (quite limited) objects more robustly and completely, on a model plainly derived from conventional humanities scholarship. That wasn't a mistake, nor are those projects without significant value. They've contributed to our understanding of, for example, Rossetti and Whitman, Stowe and Dickinson, Shakespeare and Spenser. And they've helped legitimate digital work in the eyes of suspicious colleagues by showing how far we can extend our traditional scholarship with new technologies. They've provided scholars around the world—including those outside the centers of university power—with improved access to rare materials and improved pedagogy by the same means. But we shouldn't ignore the fact that they've also often been large, expensive undertakings built on the assumption that we already know which authors and texts are the ones to which we should devote our scarce resources. And to the extent that they've succeeded, they've also reinforced the canonicity of their subjects by increasing the amount of critical attention paid to them.

What's required for computational and quantitative work—the kind of work that undermines rather than reinforces canons—is more material, less elaborately developed. The Wright collection, on which the maps are based, is a partial example of the kind of resource that's best suited to this next development in digital humanities research. It includes (or will include) every known American literary

text published in the United States between 1851 and 1875 and makes them available in machine-readable form with basic metadata. Google Books and the Hathi Trust aim for the same thing on a much larger scale, currently covering a large portion of the extant public domain and offering the hope of computational access to most of the books (in and out of copyright) held in major research libraries. None of these projects are cheap. But on a per-volume basis, they're very affordable, largely because they avoid extensive human intervention (via detailed markup, for example) in the preparation of the texts. And of course the Google and Hathi corpora were produced with remarkably limited direct costs to the academic departments that will use them, particularly in light of their magnitude. The texts they include are deeply impoverished compared to those curated in author- and subject-specific archival projects. But they're more than adequate for many of our analytical purposes, which, again, need not turn solely on close textual hermeneutics.

It will still cost a good deal to make use of these what we might call "bare" repositories. The time, money, and attention they demand will have to come from somewhere. My point, though, is that if (as seems likely) we can't pull those resources from entirely new pools outside the discipline—that is to say, if we can't just expand the discipline so as to do everything we already do, plus a great many new things—then we should be willing to make sacrifices not only in traditional or analog humanities but also in the types of first-wave digital projects that made the name and reputation of DH. This will hurt, but it will also result in categorically better, more broadly based, more inclusive, and finally more useful humanities scholarship. It will do so by giving us our first real chance to break the grip of small, arbitrarily assembled canons on our thinking about large-scale cultural production. It's an opportunity not to be missed and a chance to put our money—real and figurative—where our mouths have been for two generations. We've complained about canons for a long time. Now that we might do without them, are we willing to try? And to accept the trade-offs involved? I think we should be.

NOTES

1. How many canons are there? The answer depends on how many people need to have read a given set of materials in order to constitute a field of study. This was once more or less everyone, but then the field was also very small when that was true. My best guess is that the number is at least a hundred or more at the very lowest end—and a couple orders of magnitude more than that at the high end—which would give us a few dozen subfields in English, give or take. That strikes me as roughly accurate.

2. Wright American Fiction Project, http://www.letrs.indiana.edu/web/w/wright2.

3. Why only a third? Those are all the texts available in high-quality machine-readable format with good metadata (via the MONK Project, http://monkproject.org) at the moment.

4. Peter Warden, Geodict, http://datasciencetoolkit.org.

5. Data are also available for 1853, 1873, and 1875, each showing substantially similar features. The intervening years—1854 through 1872—have as yet too few digitized volumes for reliable use. The cluster of locations in southern Africa is a known identification error.

BIBLIOGRAPHY

Elson, David K., Nicholas Dames, and Kathleen R. McKeown. "Extracting Social Networks From Literary Fiction." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,* 138–47. Uppsala, Sweden, 2010.

Greco, Albert, Clara Rodrguez, and Robert Wharton. *The Culture and Commerce of Publishing in the 21st Century*. Stanford, Calif.: Stanford University Press, 2007.

Hope, Jonathan, and Michael Witmore. "The Hundredth Psalm to the Tune of 'Green Sleeves': Digital Approaches to Shakespeare's Language of Genre." *Shakespeare Quarterly* 61, no. 3 (2010): 357–90.

Moretti, Franco. *Maps, Graphs, Trees: Abstract Models for a Literary History*. New York: Verso, 2005.

R. R. Bowker. "New Book Titles & Editions, 2002–2009." April 14, 2010, http://www.bowker.com/index.php/book-industry-statistics.

———. "U.S. Book Production, 1993–2004." April 14, 2010, http://www.bowker.com/bookwire/decadebookproduction.html.

Wright, Lyle H. *American Fiction 1851–1875: A Contribution Toward a Bibliography*. Rev. ed. San Marino, Calif.: Huntington Library Press, 1965.