DRAFT – Final version forthcoming in *Modern Language Quarterly* (December 2017)

The Equivalence of "Close" and "Distant" Reading; Or, Towards a New Object for Data-Rich Literary History

Katherine Bode

In a recent blog post Ted Underwood (2015) describes as "malarkey" the "version of distant reading currently circulating in the public imagination" – namely, that it analyses "a massive database that includes 'everything that has been thought and said'." He continues:

> In the early days of distant reading, Franco Moretti did frame the project as a challenge to literary historians' claims about synchronic coverage (We only discuss a tiny number of books from any given period – what about all the rest?) But even in those early publications, Moretti acknowledged that we would only be able to represent "all the rest" through some kind of sample.

Underwood is correct in a narrow sense: Moretti engages in, and occasionally acknowledges the practice of, data sampling. But it does not follow that the public imagination, or the mainstream media outlets feeding it, confected the view of "distant reading" as enabling direct and objective access to a comprehensive literary historical record. Moretti's work and that of his long-time collaborator and co-founder of the Stanford Literary Lab, Matthew Jockers, provide more than ample grounds for this public perception. More particularly, while claiming direct and objective access to "everything," these authors represent and explore only a very limited proportion of the literary system, and do so in an abstract and ahistorical way.

I focus on Moretti and Jockers because they dominate academic and general discussions of data-rich literary history. Not only have they written some of the only book-length contributions to the field (Moretti 2005, 2013a; Jockers 2013a), but their work is reported on in major public forums such as the *Paris Review*, *Financial Times, New York Times*, and *New Yorker Magazine* (Piepenbring 2015; Sunyer, 2013; Schultz, 2011; Rothman, 2014; Lohr, 2013), and "distant reading" is a term now routinely applied to data-rich literary research in general (e.g. Clement 2013: para 1). Moretti's work, in particular, has also received its share of criticism, following several lines. The most resistant scholars maintain that data is inimical to literature and that only close reading can explore its nuance and complexity (e.g. Trumpener 2009). James English (2010: xii, xiii) attributes this response to the discipline's foundationally "negative relationship" to "counting," and Moretti's role in exacerbating the divide. More receptive critics advocate the use of data, but echo Moretti's (2013a: 48) account of "distant reading" as "a little pact with the devil," acknowledging that his new forms of knowledge inevitably abstract and simplify complex phenomena (e.g. Love 2010: 374). Regarding Moretti's tendency to "overestimat[e] the scientific

objectivity of his analyses" (Ross 2014), some perceive Moretti's claim to authoritative knowledge as an unfortunate side effect of his polemic against "close reading" (Burke 2011: 41), while others ascribe a more foundational essentialism to his work. John Frow (2008: 142) argues that Moretti conceives of "literary history … as an objective account of patterns and trends" by "ignor[ing] the crucial point that these morphological categories he takes as his base units are not pre-given but are constituted in an interpretive encounter by means of an interpretive decision."

In my view, these criticisms describe the symptoms – not the essence – of a problem, which in fact inheres in Moretti's and Jockers's common neglect of the activities and insights of textual scholarship: the bibliographical and editorial approaches that explore and explicate the literary historical record. In dismissing the critical and interpretive nature of these activities, and the historical insights they embody, Moretti and Jockers model and analyse literary history in reductive and ahistorical ways. Their neglect of textual scholarship is not an effect of importing data into literary history but inherited from the New Criticism: contrary to the prevailing view, "close" and "distant" reading are not opposites. Building on significant – though uneven and unacknowledged – departures from "distant reading" and "macroanalysis" by Underwood and other scholars in data-rich literary history, I present the case for a new scholarly object of analysis, modeled on the foundational technology of textual scholarship: the scholarly edition.

Those familiar with Jerome McGann's call for "philology in a new key," most recently in *A New Republic of Letters* (2014), will recognize the obvious debt my argument owes to his. In approaching "distant reading" and "macroanalysis" from this perspective, however, I seek to chart a path beyond the polemics of both sides: beyond the view of literary history as defined by either the multiplication of data points (in Moretti's and Jockers's work), or the elaboration of unique and ultimately unknowable philological objects (in McGann's). More specifically, while using the scholarly edition as a framework for data-rich literary history, my focus is not the individual literary works McGann (2014: 3) maintains as the essential basis of an "object-oriented and media approach to the study of literature and culture." Rather, I apply theoretical and practical features of the scholarly edition to core elements of Moretti's and Jockers's approach: to modeling literary systems and to the mass-digitized collections that make this historical formation amenable to analysis.

I

Underappreciated in commentary on "distant reading" and "macroanalysis" are the shifting meanings of both terms. "Distant reading" was originally proposed as a paradigm for world literary studies, not literary history, unrelated to either data or computation (Moretti 2000). With *Graphs, Maps, Trees* (Moretti 2005), literary history was foregrounded, as was data (with the "units much smaller or larger than the text," introduced in "Conjectures on World Literature" (Moretti 2000:

57), translated into data points). Computation, and the digital resources and methods it works with, were in turn central to *Distant Reading* (Moretti 2013a); but there, literary history was ceding ground to the "theory of literature" as the primary focus in the "encounter of computation and criticism" (Moretti 2013b: 9).

Although data and computation remain central, the primary object of "distant reading" is now less often literary historical systems – particular social, material, and political contexts for literary development and change – than the "concepts of literary study" (Moretti 2013b: 1). Where literary systems are comprised predominantly of the "great unread" (Cohen, cited in Moretti 2000: 55), these concepts, including characterization, plot, and dramatic form, are investigated via select, canonical literary works. Jockers's recent work demonstrates this same shift from literary history – his explicit focus in *Macroanalysis* (2013a) – to categories of literary analysis (in his case, plot).[1] Yet even as Moretti and Jockers have moved from an historical to a conceptual emphasis, "distant" reading" and "macroanalysis" dominate – and limit – public, and much academic, perception of what data-rich literary history entails.

Pace Underwood's defense, both Moretti and Jockers (albeit to different extents) present data and computation as providing direct and comprehensive access to the literary historical record. In Moretti's early historical work, data alone served this purpose, with repeated references to literary data as "facts," "ideally independent of interpretations"; "*data*, not interpretation"; and "useful because they are independent of interpretation" throughout the first chapter of *Graphs* (2005: 3, 9, 30). On this basis, Moretti accorded his claims an unrealistic exactitude – for instance, asserting that bibliographical data "can tell us when Britain produced one new novel per month or week or day, or hour for that matter" (9) – and presented data visualization as a transparent window onto history: "graphs, maps, and trees place the literary field literally in front of our eyes – and show us how little we still know about it" (2). The same construction of literary data as factual and transparent appears in *Distant Reading*, where Moretti (2013a: 211) celebrates data visualization as providing "a set of two-dimensional signs … that can be grasped at a single glance." Such descriptions, which substitute seeing what is there for the interpretive acts involved in constructing literary data, organizing it, and ascribing an historical explanation to the results, underpin Moretti's (67) claim to explore "the literary field as a whole."

Missing in the extant critiques of such claims is their underlying cause: Moretti's disinterest in the scholarly infrastructure underpinning his arguments. Where the data come from analog bibliographies, as in the first chapter of *Graphs* and the study of 7000 titles in *Distant Reading*, parentheses and footnotes occasionally admit that comprehensive access to the facts of literary history is not achieved. For example, Figure 7 in this latter study – showing the number of British novels – stops in 1836 (where the other graphs extend to 1850), and a footnote (2013a: 188)

comments, "it seems very likely that Andrew Block's bibliography significantly overstates the number of novels published after that date." Yet acknowledging that a particular set of literary data is the outcome of a ("significantly" flawed) "interpretive encounter" affects neither Moretti's rhetoric or subsequent analysis: the chapter still claims to "read the entire volume of the literary past" (58), and while the data is absent from Figure 7, Block's bibliography is the only source for titles published from 1836 to 1850. Moretti proceeds, in other words, analyzing titles he knows never existed.

Where literary data derived from analog representations of the literary historical record are only "ideally independent of interpretation," Moretti regards mass-digitized collections as actually independent.[2] With collections becoming the rhetorical, if not the primary analytical, focus of *Distant Reading,* Moretti (181) looks forward just "a few years," to when "we'll be able to search just about all novels that have ever been published and look for patterns among billions of sentences," and notes that, where literary studies has previously experienced "the rise of quantitative evidence … without producing lasting effects, … this time is probably going to be different, because this time we have digital databases and automatic data retrieval." The disregard for the specifics of the disciplinary infrastructure manifested by these claims was reinforced in a recent interview where Moretti (2013c) aligns digital humanities with three elements:

> new, much larger archives; new, much faster research tools; and a (possible) new explanatory framework. The archives and the tools are there to stay; they are important but not intellectually exciting. What appeals to me is the prospect of a new explanatory model – a new theory and history of literature.

Rather than "there to stay," digital archives, like bibliographies, are interpretative constructs; and they are still evolving, not only in content but in form, in the process presenting significant practical and conceptual challenges for literary history.

The assertion of comprehensive access to the literary historical record is even more essential to Jockers's "macroanalysis." Foundational to this approach is Jockers's (2013a: 6) view that any form of interpretation is defective: "interpretation is fuelled by observation, and as a method of evidence gathering, observation – both in the sciences and in the humanities – is flawed." Where interpretation and observation are "anecdotal and speculative" (31), "big data" is supposedly separate from human involvement and thus offers "comprehensive and definitive" historical facts. According to Jockers (7–8), literary scholars "have the equivalent of big data in the form of big [digital] libraries … [or] massive digital-text collections," that enable "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." As Jockers (20) says of one of Moretti's analyses, this unprecedented and uninterrupted access to the matter of literary history "leaves little room for debate." Jockers employs scientific metaphors to buttress this

association of scale and comprehensive access, the most explicit being "open-pit mining or hydraulicking" (9–10). While "microanalysis" (including reading and digital searching) discovers "nuggets," "macroanalysis" accesses "the deeper veins [that] lie buried beneath the mass of gravel layered above." Working with the "gravel" of literary history enables him "to unearth, for the first time, what these corpora really contain," a metaphor that conflates analysis with achieving total access.

Where Moretti occasionally acknowledges limitations in his data (before proceeding with analyses regardless), Jockers (28) maintains the view that any "leap from the specific to the general" is flawed until the book's final chapter. There he admits to the obvious gap between his datasets and the "population" of nineteenth-century novels, describing his largest "corpus of 3,346 texts" as "incomplete, interrupted, haphazard." This acknowledgement that the "comprehensive work is still to be done" generates an awkward comparison of "macroanalysis" with Charles Darwin's theory of evolution. Where both are "idea[s]" – because "there are further dimensions to explore" – literary scholars are advantaged over evolutionary biologists "in terms of the availability of our source material" (171–2). In a context where bigger is better – as Jockers (25) puts it elsewhere in the book, "eight is better than one, [but] eight is not eight thousand, and thus, the study is comparatively anecdotal in nature" – his "3,346 observations and 2,032,248 data points" (172) are seemingly indicative of knowledge in and of themselves. Jockers (175) concludes the book by admitting one impediment to "macroanalysis," but it is only legal: though almost "everything has been digitized," post-1923 publications remained (at the time of writing) protected by copyright, leaving literary scholars dependent on legal reform before they might realize "what can be done with a large corpus of texts".[3]

Although Jockers is no longer part of the Stanford Literary Lab, and while it is not clear what part Moretti played in the project,[4] the most recent collaboration from that group departs, in one important way, from the approach to literary history, data, mass-digitization, and computation I have described. Pamphlet 11 pays detailed attention to the gaps between "the published" (all literary works in history), "the archive" (the portion of what was published that has been preserved and is now increasingly digitized), and "the corpus" (the segment of the archive selected for a particular research question). Although imagining that the "convergence of these three layers into one … may soon be reality" (Algee-Hewitt, et al. 2016: 2), the authors acknowledge the current impossibility of achieving this state, and the constructed – and selective – nature of literary data. Despite this distinction, Pamphlet 11 follows Moretti's and Jockers's precedent in misconstruing the nature of our disciplinary infrastructure. In presuming to overcome the selections and biases of mass-digitized collections by using analog bibliographies as the basis for generating "a random sample" of what was published, the authors overlook the fact that both are derived from "the archive,"

predominantly the collections of the major university libraries. Pamphlet 11 also replicates Moretti's and Jockers's approach in not publishing its datasets.[5]

Moretti often references his sources of data – chapter one of *Graphs*, for instance, begins by listing the bibliographies it draws upon (2005: 5) – and advocates data sharing: "because data are ideally independent from any individual researcher, [they] can thus be shared by others, and combined in more ways than one." However, he does not enact this practice. Jockers occasionally publishes the results of data analysis – such as the 500 themes developed from topic modeling (presented as word clouds on his website) (Jockers 2012) – but does not provide the textual data analyzed, even at the level of word frequencies,[6] and is significantly less open than Moretti about the source and composition of his datasets. Indeed, I have discovered only one instance where Jockers indicates the titles and authors he investigates: and then, only 106 of the total 3,346 (identified in the context of reporting confusion matrices) (Jockers 2013b).

In Moretti's case, one might suppose it possible to reconstruct his datasets from cited sources. But his account (in an appendix to *Graphs*) of creating the dataset for "British novelistic genres, 1740-1900," highlights why this is not feasible. There he describes his periodization as "not always explicit" (31) in the bibliographies, thus evincing the role of his – unpublished and thereby unspecified – interpretive decisions in data construction. Even if Jockers listed the titles and authors he analyzed, it would be impossible to reconstruct the basis of his arguments without access to the textual corpora he uses (which are not just texts of literary works, but highly prepared – or pre-interpreted – selections from those texts). Far from an incidental oversight, not publishing maintains the fiction that literary data are prior to interpretation: it removes the need either to describe the procedures for collecting, cleaning, curating, or more generally constructing their datasets, or to expose the inevitably selective and limited collections that result from that construction.

The meaning derived from a literary historical dataset – like the interpretation of an individual literary work – is shaped, profoundly, by the theoretical framework through which it is approached, and the selections and amplifications that framework produces. Accordingly, two scholars can read the same dataset – like the same literary work – and derive different meanings. Where an independent observer may be more or less convinced by the different arguments, deciding between them depends upon access to the object on which they are based. In the absence of data publication, "distant reading" and "macroanalysis" are analogous to a literary scholar finding a set of documents in an archive or archives and transcribing them, analyzing those transcriptions, and publishing the findings as demonstrating a new and "definitive" perspective on the literary field, without enabling anyone to read the transcriptions (or in Jockers's case, without revealing the titles of most of the original documents).

II

Central to Moretti's and Jockers's approach to literary history is a conception of literature as "a collective system that should be grasped as such" (Moretti 2005: 4).[7] Although the approach has significant antecedents, not least in book history,[8] this foregrounding of data-rich models of literary systems as a primary unit of analysis has influenced literary history profoundly, as demonstrated by the extensive debate about the role and relationship of reading and data in "distant reading." Yet in not recognizing the critical and constructive nature of the scholarly infrastructure they use, both authors ultimately fail to capture the historical nature of literary works, and how they connect to produce literary systems. In many cases, the results of such analyses are tautological, with the same basic data providing both the problem and the explanation.

Moretti and Jockers typically locate literary works according to the date of first book publication and the author's nationality, constituting them as a literary system when they share these basic features: as in "nineteenth-century" or "British" novels. This basic understanding of a literary system is evident, for instance: in Moretti's analyses of "7,000 titles (British novels, 1740 to 1830)" (2013a: 179–210) or of eighteenth- and nineteenth-century British novels defined in terms of the authors' gender, or subgenres of fiction (2005: 26-27; 28-30); and in Jockers's (2013a: 37) exploration of "758 works of Irish-American prose literature spanning 250 years," or 106, or 3,346 nineteenth-century British and American novels.

Depending on the reliability of the source, such datasets can enable key insights into new literary production. Jockers's study of Irish-American prose pursues an approach manifested in other digital book historical projects – some of my work included[9] – of using publication data to test existing literary historiographical arguments. Employing a dataset with the date of first publication, as well as "the geographic settings of the works, author gender, birthplace, age, and place of residence" (36), Jockers challenges Charles Fanning's claim that there was a "lost generation" of Irish-American authors for the period from 1900 to 1930, and proposes a likely explanation for this misperception: a predominance of eastern male authors in the canon – and hence, in critical assessments – of Irish-American literature (38-48). Moretti's work on new literary production extends beyond testing and revising particular literary historiographical claims, and is highly innovative in this respect. His study of titles, for instance, investigates a category of literary data that has not, as far as I know, been subject to synoptic, stylistic analysis previously (Moretti 2013a: 179–210); more broadly, Moretti combines multiple bibliographies to explore relationships between different literary historiographical claims (as in his discussion of new British novel genres or gender trends in British novel publication) (Moretti 2005: 26–29, 17–20).

But literary works are not defined by a single time and place. William St Clair's *The Reading Nation in the Romantic Period* (2004) aptly diagnoses the limitations of this approach.

Like Moretti and Jockers, St Clair (2) rejects what he calls the "parade of authors" convention in literary history, where canonical authors file past the commentator's box in chronological order, taken as representative of the historical period in which they wrote. But he equally dismisses the "parliament of texts" approach, where literary works first published at a particular time, and often, by authors of a particular nationality, are understood as "debating and negotiating with one another in a kind of open parliament with all the members participating and listening." Literary systems, after all, may include "texts written or compiled long ago and far away" (3), and some literary works are more widely published, circulated, read, and referenced than others.

New domestic literary production, the focus of Moretti's and Jockers's studies, is only a subsection of the literature in circulation at any time and place. The date of first book publication overlooks the differing availability of literary works in the years after they are published; and the first book edition is not necessarily – for many periods is rarely – the first time a literary work is available. Some titles are never published as books, and many literary works – whether in book or other formats – are republished, sometimes on multiple occasions. Likewise, an author's nationality is a poor marker for the geographical existence of a literary work in a marketplace that has been globalized since at least the eighteenth-century.

Even in circumstances where new book publications are the primary focus, this approach to modeling literary systems ignores most differences between literary works, and hence, most dynamics of those systems. Reflecting on Moretti's work, David Brewer (2011–12: 162) notes that flattening the literary field "undoes the monumentalizing that so often accompanies the literary canon," but at the expense of ignoring the varied profiles and presences of literary works in history. The popularity of different literary works at the time they are published and in subsequent generations accords them "a massively different footprint" in history, altering their influence and hence their meaning. But commercial success is not the only relevant factor. As emphasized in textual scholarship on the material and social dimensions of literature, multiple issues shape the meaning that accrues to literary works, extending from the documentary forms they take, to the relative positions and prestige of the individuals and institutions involved in producing them (authors, publishers, editors, illustrators, booksellers, advertisers) and the interconnected systems (economic, religious, educational, legal, geopolitical) in which they circulate.[10]

Where textual scholars accordingly conceptualize literary works as events, unfolding over time and space and gaining different meanings in the connections thereby formed,[11] Moretti and Jockers construct literary systems as comprised of singular and stable entities, while also imagining that they capture the complexity of such systems in the process. In fact, because their datasets miss most historical relationships between literary works, their analyses are forced to rely upon basic features of new literary production to constitute both the literary phenomenon requiring

explanation, and the explanation for it. *Macroanalysis* purports to investigate "the context in which [literary] change occurs" (2013a: 156), chiefly by analyzing words in nineteenth-century novels. What Jockers actually shows is the capacity of his computational method (a combination of stylistic analysis, topic modeling, and network analysis) to predict whether a particular work (or rather, a "bag of words" from that work) was by a man or woman, from a particular decade, of a particular genre, or by an author of a particular nationality, from a corpus defined according to those parameters. Notwithstanding the variable accuracy of this approach for different categories,[12] the methodological demonstration is impressive for extending stylistic analysis beyond small groups of documents.

But the methodological achievement does not translate into historical insight because the study considers only an abstract amalgam of literary works. In reducing context to a small number of predetermined categories, Jockers is confined to stating their presence. He cannot offer any alternative influences, nor comment on the extent to which gender, nationality, and chronology shape literary history (except implicitly, in the proportions of titles misidentified by his models). The approach yields very general, and I would argue, self-evident statements. To give examples drawn from the conclusions to Jockers's various chapters: "the linguistic choices an author makes are, in some notable ways, dependent upon, or entailed by, their genre choices" (104); "there are both national tendencies and extranational trends in the usage of … word clusters" (114); "a writer's creativity is tempered and influenced by the past and the present" (156); and "thematic and stylistic change does occur over time" (164). The generality of these conclusions is predetermined by the dematerialized and depopulated conception of influence underpinning the analysis: the model constitutes literary works as a system based on the date of (presumably first book) publication. Any book included in the system is understood to influence the system in a chronologically discrete manner, in disregard of the actual conduits of literary influence (which require availability to readers who buy, borrow, and sometimes write, literary works). Because he is modeling a diffused and generalized system, the "influence" of gender, genre, temporality, and nationality is in turn diffused and generalized.

The inadequacy of this conception of literary systems is foregrounded when Moretti considers readers, who, as Anne DeWitt (2015: 162) notes, "are both central to his argument and absent from his evidence." Lacking such information, Moretti takes literary data on publication and/or formal features of literary works as both expressive of, and explainable by, the actions of readers and the market. We can see this strategy in Moretti's (2005: 5) discussion of the first graph in *Graphs*: the "rise of the novel" across a number of national contexts (Britain, Japan, Italy, Spain, and Nigeria) at different times. Leaving aside the question of whether his graph depicts the numbers he attributes to it,[13] Moretti ascribes the leap "from five-ten new titles per year … to one new novel

per week" to "the horizon of novel-reading," the shift in the market that occurs when the novel is transformed from "an unreliable commodity" to a "*regular novelty*: the unexpected that is produced with such efficiency and punctuality that readers become unable to do without it." The argument makes intuitive sense, but it presumes that only – and all – new titles by authors of particular nations were available to – and read by – readers of those nations.

A similarly circuitous mode of argumentation characterizes "The Slaughterhouse of Literature" chapter in *Distant Reading*. Moretti (2013a: 67) proposes a framework for canon formation, wherein readers are the "butchers" of literary history "who read novel A (but not B, C, D, E, F, G, H …) and so keep A "alive" into the next generation, when other readers may keep it alive into the following one, and so on until eventually A becomes canonized." Nominating formal choices as the reason readers select particular titles over others, Moretti identifies the presence of decodable clues in Arthur Conan Doyle's fiction as the reason that author was selected by generations of readers to attain his now canonical status. Again, this is an interesting, but circular, argument. Moretti acknowledges one of the ways in which his claims are "tautological": "if we search the archive for one device only … all we will find are inferior versions of the device, *because that's really all we are looking for*" (87, original italics).

Yet the same problem – of assuming the shape of the past from that of the present – occurs at a larger scale, in that Moretti takes as transparently true the idea that authors who have a canonical status in the present were selected from the time of first publication. This argument is intrinsic to his evolutionary model, and while Moretti (68) supports it by citing an empirical study, others falsify it: St Clair (2004), for instance, demonstrates the minute readerships of five of the "big six" Romantic male poets (excepting Byron) that form our contemporary canon for that period. Whereas, in the earlier study, Moretti aligned publication with reading (a title was published, ergo it was read), in this instance Moretti's argument requires titles to be published but not read. What determines if titles were read is whether they had decodable clues; thus, once again, a feature of the data (the presence or absence of decodable clues) is used both to indicate and explain the activities of readers.

Moretti (2005: 1) has said, in defense of his method, that reducing literary works to one or two features is part of the "specific form of knowledge" "distant reading" provides: "fewer elements, hence a sharper sense of their overall interconnection. Shapes, relations, structures. Forms. Models." My argument is not against reduction and abstraction per se. While particularly obvious in data-rich studies (because reliant on identifying attributes that can be represented in uniform fields), reduction and abstraction characterize all analysis (even so-called "close readings" do not interpret literary works as a whole but particular, extracted instances of particular, abstracted features of those works). But in misapprehending the historical nature of literary works, the

assumptions Moretti and Jockers make in modeling literary systems lose any useful sense precisely of "interconnection": how literary works exist and relate to one another in socially, spatially, and temporally relevant ways, not in terms of potentially, and certainly relatively, abstract categories of production.

III

Where "distant reading" and "macroanalysis" are celebrated – or decried – for their departure from "close reading," these approaches share the same disregard for textual scholarship, and assumption that literary works are stable and singular entities. "Close reading" is largely protected from the worst consequences of these underlying assumptions by the documentary and infrastructural context in which it occurs. But the same cannot be said of "distant reading" and "macroanalysis," whose arguments take the core premise of the New Criticism – that the text is the source of all meaning – to an extreme conclusion. Because the rise of quantitative evidence in literary history is perceived as the entry of a foreign paradigm, current responses consider the possibilities – or problems – of a new form of analysis. Yet such responses fail to recognize that analysis is inevitably a function of the object analyzed. What literary history needs is not "close" or "distant" reading, nor a simple integration of the two, but a new scholarly object for representing literary works in their historical context, one capable of managing the documentary record's complexity, especially as it is manifested in emerging digital knowledge infrastructure. Building on existing work in digital humanities, I adapt the theory and practice of the scholarly edition to mass-digitized collections, and the modeling of literary systems on that basis.

As is well known, the New Criticism, and its core method of "close reading," was an early- to mid-twentieth century literary movement that rejected the historical – biographical, material, sociological – concerns of previous scholarship to privilege the text as the source of all meaning. The subsequent critique of this movement is also well established; and the focus on both text and context in many subsequent forms of literary history – feminism, postcolonialism, new historicism – explicitly rejected the New Critical view of the text as a self-contained and self-referential aesthetic object. Despite the apparent demise of the New Criticism, the continuing centrality of "close reading," and its rhetorical focus on "the text," perpetuates the earlier movement's dismissal of textual scholarship (Cain 1982). As McGann (2004) and others have noted (e.g. Eggert 2009), the assumption that literary works are texts, and texts are single, stable, and self-evident entities, dismisses the documentary record's multiplicity, and with it, the critical contributions of those disciplines (particularly bibliography and scholarly editing) dedicated to investigating that multiplicity.

The marginality of textual scholarship – not the ascendancy of quantitative evidence – underpins the problems I have described in Moretti's and Jockers's work. In turn, their arguments mirror the New Criticism's perception of the text as the source of all meaning, even as the "texts" under consideration have expanded from particular versions of literary works to particular versions of bibliographical and textual data. Moretti's discussion of readers, based only on publication data, manifests the view that "the text" contains all that is relevant to interpreting it. In treating the literary system as a dispersed linguistic field, Jockers takes this rhetoric to its extreme conclusion, proposing a literary historical world where there are no structures beyond the textual. "Signals" of gender, genre, or nationality, comprised entirely of word frequencies, are substituted for gender, genre, or nationality as historical and cultural constructs. Far from the opposite of "close reading," the dematerialized and depopulated understanding of literature in Jockers's work enacts the New Criticism's neglect of context, in a manner rendered only more abstract by the large number of "texts" under consideration.[14]

For the most part, "close readings" are protected from the abstraction inherent in the rhetoric of text by the knowledge infrastructure in which they are embedded, and by a focus on specific documents. Scholarly editions provide critics with carefully historicized texts for consideration; when one is not available, the standard practice of bibliographical referencing ties discussion of "the text" to a particular version of the work; and because close readings analyze particular versions, any discussion of "the text" is contextualized by the information about the work's history contained in the material form the critic assesses. Yet even with these protections, the centrality and assumed singularity of the text – and the disassociation from the literary work's complex historical existence this produces – can negatively impact the capacity of "close reading" to investigate literary history. For example, discussing Wilkie Collins's *Moonstone*, Mary Elizabeth Leighton and Lisa Surridge (2009: 207) note the varied interpretations critics offer of its meaning for nineteenth-century readers: from tale of "imperialist panic" to class critique. While all believe "they are reading the same text," in fact the work "took on strikingly different forms – and hence different meanings – in different markets," including British and American serializations. Assuming that the modern document they read is identical to what circulated in the past, these critics project textual singularity onto an historical context characterized by documentary multiplicity.

A broader way of expressing the point is to say that outcomes of analysis are inevitably tied to the object analyzed. When a gap exists between the contemporary object assessed and the historical object it supposedly represents – and when the critic is unconscious or dismissive of that gap – no degree of nuance or care in the reading can supply that historical meaning. Here we come to the fundamental problem with the most common, now almost reflexive, response to the rise of quantitative evidence in literary history. Ever since Moretti proposed "distant reading," multiple

calls have been made – including by Moretti and Jockers – to integrate "traditional and computational approaches" (Gibbs and Cohen, 2011: 70; Moretti 2013a: 181; Jockers 2013a: 26). Understood in terms of the different perspectives the two offer literary history, this strategy is eminently sensible: data-rich analysis has the potential to explore large-scale patterns and connections in ways that non-data-rich research cannot; likewise, traditional textual analysis can provide insights into the meaning of particular literary works that quantitative studies cannot. But the focus on methodological capacities or limitations overlooks the lack of an appropriately historicized object for data-rich analysis, and more specifically, the fact that producing such an object is itself a critical and interpretive enterprise.[15] Lack of a scholarly object capable of representing literary historical systems – that is, literary works that circulated and generated meaning together at particular times and places – is the real reason it has proven so difficult, in practice if not in theory, to integrate "traditional and computational methods" for literary historical investigation.

Fortunately, beyond the work of Moretti and Jockers, and of those who adopt their methods, researchers are moving toward supplying this lack. In digital literary studies broadly, an explicitly editorial and bibliographical approach to working with digital collections is apparent, including in research at the intersection of digitized collections and literary systems. Whether building digital collections that propose some historical relationship between literary works and/or authors,[16] or using mass-digitized collections to historicize particular titles and authors,[17] a number of projects attend closely to different manifestations of literary works, as well as the partiality of the digitized record, and strategies for accommodating it. For research that engages with literary systems directly, in the manner I have been describing – by constructing and analyzing data-rich models of these conceptual entities – this editorial and bibliographical consciousness is most consistently apparent in an emphasis on the historical existence of, and connections between, literary works.

Where Moretti and Jockers assume that works first published around the same time and by authors of the same nation automatically constitute a literary system, other projects model literary systems in terms of specific spatial, temporal, and social interconnections between literary works. Richard So and Hoyt Long (2013: 148) explore "the collaborative networks that underwrote the evolution of modernist poetry" based on the poets published in the same American, Japanese and Chinese periodicals; Anne DeWitt (2015) study genre formation based on titles mentioned in the same reviews. The "Viral Texts" project identifies reprinted passages in historical newspapers and magazines where popularity, as well as personal and structural connections between newspaper editors, is indexed by republication (Smith et al. 2015; Cordell 2015).

All of these projects recognize that literary works do not exist in a single time and place, but accrue meaning in the contexts in which they are produced and received. Some – including the

"Viral Texts" project – build on this foundation by providing detailed accounts of the construction of their datasets, and publishing them. In their article on literary prestige in nineteenth- and early-twentieth century America and Britain, Ted Underwood and Jordan Sellers (2015) explain that the project's most time-consuming element was not constructing and training their model – which characterizes literary works in terms of whether they were reviewed or not – but identifying the different subgenres – poetry, prose fiction, and drama – in the HathiTrust Digital Library. Not only do Underwood and Sellers publish the datasets and code underpinning their work, but in collaboration with HathiTrust, Underwood (2015) has published word counts by subgenre, with yearly summaries.

While the departure from Moretti's and Jockers's approach is significant, these other projects are yet to confront – at all in some cases, fully in others – the challenge posed by the contingency of our disciplinary infrastructure, and of the datasets derived from them: what the authors of the recent Stanford Literary Lab pamphlet describe as the difference between "the published," "the archive," and "the corpus" (Algee-Hewitt, et al. 2016). Most of the projects discussed above ignore these gaps, or note their existence without examining their scope or potential effects.[18] Underwood and Sellers recognize the difference between the historical context investigated and the digital collection used in that process, and consider its impact on specific findings. For instance, knowing that HathiTrust "mainly aggregates the collections of large American libraries" enables them to explain why their model "makes more accurate predictions" for American poetry collections: their sample over-represents such titles (2015: 22–23).

Yet even this project is equivocal in characterizing the broader relationship between historical literary systems and our disciplinary infrastructure, and the implications of this situation for data-rich literary history. In a footnote, Underwood and Sellers (2015: 5, fn11) note that HathiTrust "may represent more than half of the titles that were printed" because it contains "about 58% of titles recorded in standard bibliographies," while also finding that HathiTrust contains "many titles left out of" existing bibliographies. This observation identifies a significant lack of overlap between established bibliographical records and the holdings of this major, mass-digitized collection; but Underwood and Sellers do not discuss the implications of this situation, either for their study or for data-rich literary history broadly.

Ultimately, neither the analog nor the digital record offer an unmediated and comprehensive view of the literary historical record; both are partial, and not necessarily in complementary ways. The nature of the challenge thereby facing data-rich literary history – of proposing an historically coherent whole (a literary system) from a collection, or collections, of parts (the disciplinary infrastructure and the literary data and digitized documents it presents) – also suggests the means of meeting it. The scholarly edition has long offered textual scholars both a theoretical foundation and

a practical mechanism for demonstrating and accommodating the documentary record's contingency and partiality. Applied to the literary system rather than the literary work, the scholarly edition offers a framework for acknowledging that models of literary systems are not simply arguments about the existence of literary works in the past; they are arguments made with reference to the bibliographies and collections, analog and digital, that transmit historical evidence of those works and their interrelationships, and the interpretive processes that translate that evidence into literary data. A scholarly edition of a literary system offers a format capable of investigating that history of transmission and its effects, publishing the results of that process, and through this dual function, extending the insights of such arguments to the discipline of literary history as a whole.

In making this claim, I am not proposing that a scholarly edition of a literary work and of a literary system should, or ever would, be equivalent. While neither exists "in fact," the literary work is a "regulative idea" (Eggert 2013: 9) shored up by extensive social, economic, institutional, and technological structures, whereas the literary system is a loosely defined conceptual entity of relevance primarily to literary scholars. Nor am I suggesting that data-rich literary historians perform the activities associated with scholarly editing of literary works: for instance, comparing existing versions to provide a reliable critical edition. Rather, the scholarly edition's potential as the basis for data-rich literary history lies in the simultaneously theoretical and practical approach to the documentary record that underpins it.

Far from simply a version of a literary work, a scholarly edition is an "embodied argument about textual transmission," as Paul Eggert (2013: 177) puts it; or, in McGann's (2005: 203) words, a "hypothetical platform" for historical enquiry. Elaborating, McGann argues that a scholarly edition "discloses the hypothetical character of its materials and their component parts as well as the relationships one discerns among these things," and thus both demonstrates and provides a pathway through the instability of the "textual condition".[19] The dual function of the scholarly edition inheres in the interrelationship between its two essential parts: the critical apparatus and reading text. Where the former describes and justifies the scholarly editor's engagement with the documentary record constitutive of the literary work, and specifically, with the inevitable gaps and uncertainties that engagement exposes, the latter represents the outcome of the extended critical encounter, offering to those who accept its tenets – or lack the expertise to engage with the literary historical record in this manner – a basis for analysis.

For the scholarly edition of a literary system I am proposing, the critical apparatus describes the history of transmission constitutive of the literary system modeled. Much more than simply recounting the construction of a dataset – something already offered (though by no means universally) by existing projects in data-rich literary history – as much as possible such a critical apparatus outlines the complex relationships between the historical context explored; the

disciplinary infrastructure employed in understanding that context; the decisions and selections implicated in creating and remediating that collection;[20] and the additional transformations wrought by the editor's extraction, construction, and analysis of that data. In providing this history of transmission, the critical apparatus elaborates, with respect to the modeled literary system, the complex strands of production and reception – including the current moment of interpretation – McGann (2014: 82) identifies as the "double helix" producing all cultural objects. Where "the stylistic protocols of literary criticism" mean that issues deemed methodological are often relegated to footnotes or "methodological caveats" (Underwood and Sellers 2015) – as if they qualified rather than constituted the basis of the arguments offered – the critical apparatus provides a dedicated site for demonstrating and justifying the foundational argument of data-rich literary history: its modeling of a literary system.

In time, data-rich literary history might develop formats and abbreviations, similar to those used in critical apparatuses for scholarly editions of literary works, to describe common issues affecting the sequence of connections (and disconnections) between the historical context investigated and the literary data employed for that purpose. For the scholarly edition of a literary system I am building – of fiction serialized in nineteenth-century Australian newspapers – my critical apparatus is simply an essay, tracing a history of transmission from the context in which these newspapers were originally produced and received; to their collection and remediation, ultimately as digitized documents in the largest collection of historical newspapers internationally (the Newspapers Zone of the National Library of Australia's Trove Database); to my representation of those documents as textual and bibliographical data. This account focuses on three issues I see as imperative to justifying the reliability of my model for historical analysis: the representativeness of the newspapers digitized; the consistency of my method in identifying fiction in these digitized documents; and the claims about production and reception presented by my data model.[21] However, as in a scholarly edition of a literary work, the issues foregrounded by the critical apparatus will reflect what the editor (or editorial team) perceives as most relevant to understanding that literary system and its relationship to the disciplinary infrastructure that evidences it .

The reading text manifests – demonstrates and, more particularly, publishes – the outcome of this careful exploration of the successive acts of production and interpretation ultimately constitutive of the modeled literary system. In the form of literary data – preferably both bibliographical metadata and digitized texts – it provides a stable and accessible representation of an historical literary system for others to investigate, for either traditional literary historical or data-rich research. In its stability, such a reading text resembles the curated collection of data Underwood has published in conjunction with HathiTrust. Such stability is vital: where all documents and collections are "transactional" (Drucker 2014: 12) entities – effects of instances of

production and reception – the multiple, interactive processes constitutive of digital documents and collections make this situation more acute. In publishing the outcome of a particular, critical encounter with digital disciplinary infrastructure, both the Underwood/HathiTrust dataset and the reading text for a scholarly edition of a literary system offer a consistent object for analysis that does not presume or pretend that the historical record exists in any stable form, independently of the structures and systems through which we access and assess it.

At the same time, the reading text I am proposing differs in two important respects from the word counts offered by Underwood/HathiTrust. First, and most basically, where the latter is only usable by those with programming expertise, or access to it, the reading text of a scholarly edition of a literary system is accessible to all literary scholars, and as such, requires an interface for searching and browsing literary data as well as facilities for export. Second, while the Underwood/HathiTrust collection offers raw data from which researchers are invited to construct "the sample you need for your research",[22] the reading text of a scholarly edition already embodies an argument about what constitutes an historically relevant and justifiable sample for analysis. Obviously, such a reading text cannot be used to explore all literary historical contexts, and in this sense, it is less potentially extensible than the Underwood/HathiTrust collection (although that collection, also, is not applicable to all literary historical research, given that it relates to a specific – albeit extensive – time period and collates a particular type of text: predominantly English-language American editions). The reading text for a scholarly edition of a literary system is not, however, intended as raw data but as an argument about the historical existence of a collection of literary works, and an interpretive intermediary between increasingly complex, digital disciplinary infrastructure and the requirements of literary historical analysis.

My reading text, when published, will present detailed bibliographical metadata and digitized text for approximately 10,000 stories serialized in nineteenth-century Australian newspapers, many of which are not known to have appeared in Australia; some of which are not known to have been published at all. It will offer this data in a form accessible to all literary historians (as a searchable and browsable database, and as a selectively – or completely – downloadable dataset). As with a traditional scholarly edition, the reading text's composition reflects what the editor (or editorial team) perceives as most relevant to understanding a particular documentary context. For instance, I see the permutations that literary works underwent as they were circulated and republished in nineteenth-century Australian newspapers as vital to understanding both Australian and global literary culture of the period; accordingly, I provide both standard bibliographical details as well as information relevant to the publication event (for instance, changes in the title or the designation of authorship as anonymous).

For some researchers, such a reading text will function in much the same way as existing digital collections, providing a site for searching or browsing the digitized documentary record. In this capacity, however, it enables publication of what could (and should) be a key contribution of data-rich literary history to the broader field: a massively expanded bibliographical record (indicated by the, who knows how many, new works not previously known to literary historians uncovered in Underwood and Sellers's exploration of the HathiTrust Digital Library). For other researchers, including those currently using mass-digitized collections to locate particular authors and works in the historical context in which they operated, the reading text for a scholarly edition of a literary system will provide a carefully, consistently, and – by the critical apparatus, explicitly – historicized digital collection for this task; and for researchers interested in analyzing large-scale trends in the publication, circulation, and reception of literary works, the reading text would offer a "shared" dataset, that could be analyzed and "combined in more ways that one" (Moretti 2005: 5).

Grounding data-rich literary history in scholarly editions of literary systems emphasizes that constructing literary data is just as much an interpretive and critical activity as its analysis; and that the historical nuance of such analyses foundationally depends on the historical knowledge embedded in those constructions. With the scholarly edition of a literary system already, in and of itself, an argument about the "collective system … as a whole" (Moretti 2005: 4), analyses of its reading text can attend to its various features and dimensions. Depending on interest and inclination, literary historians can ask of the model – as Alan Liu (2008) does in analogizing postmodern historicism with the act of querying a database – how does this complex system I am investigating appear from this particular perspective? Like a scholarly edition of a literary work, a scholarly edition of literary system is thus intended not to conclude, but to enable various forms of, investigation, including those that move between the single literary work and the system in which it existed and operated.

The approach I have advocated does not take the path increasingly recommended for data-rich literary history: of integrating scientific and social scientific measures of statistical uncertainty into historical analysis (e.g. Goldstone 2015). Given that the construction of literary data is an historical argument made in the context of a history of transmission – the effects of which are difficult to qualify, let alone quantify – I do not see that any assessment of error is made more useful or concise by its numerical expression. Instead, in the intersection of its critical apparatus and reading text, a scholarly edition of a literary system offers an explicitly theorized and practical framework for exploring the multiple instances of production and reception constitutive of the data-rich models analyzed, and assessing and managing the inevitable contingency of that situation for historical analysis.

As the basis for data-rich literary history, a scholarly edition of a literary system brings promising trajectories to fruition, while overcoming limitations in the work of the field's two most high-profile practitioners: Moretti and Jockers, who claim to represent everything – directly, comprehensively, and objectively – while exploring only a limited proportion of the literary system. As I have argued, these problems are not the result of integrating data into literary history, nor do they exist in opposition to "close reading." Rather, they occur because "distant reading" and "macroanalysis" adopt and perpetuate the disregard for textual scholarship foundational to the New Criticism, without benefiting from the institutional and infrastructural protections afforded to its core method. A scholarly edition of a literary system supplies such supports and constraints for data-rich literary history, while extending the insights gained from that field's engagement with emerging digital infrastructure to the discipline as a whole. Conducted on that basis, data-rich literary history could transform from an unexpected, often unwelcome, intruder into a vital interlocutor between the discipline as a whole and the digital context in which it increasingly operates.

BIBLIOGRAPHY

Alfano, Veronica, and Andrew Stauffer, eds. 2015. *Virtual Victorians: Networks, Connections, Technologies.* Houndsmill and New York: Palgrave Macmillan.

Algee-Hewitt, Mark, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser. 2016. "Canon/Archive. Large-scale Dynamics in the Literary Field." *Literary Lab Pamphlet 11*, http://litlab.stanford.edu/LiteraryLabPamphlet11.pdf

Allison, Sarah, Ryan Heuser, Matthew Jockers, Franco Moretti, and Michael Witmore. 2014. "Quantitative Formalism: An Experiment." *Literary Lab Pamphlet 1*. http://litlab.stanford.edu/LiteraryLabPamphlet1.pdf.

Bode, Katherine. 2012. *Reading by Numbers: Recalibrating the Literary Field*. London: Anthem Press.

Bode, Katherine, and Carol Hetherington. 2014. "Retrieving a World of Fiction: Building an Index – and an Archive – of Serialized Novels in Australian Newspapers, 1850–1914." *Script and Print* 38, no. 4: 197–211.

Brewer, David A. 2011-2012. "Counting, Resonance, and Form: A Speculative Manifesto (with Notes)." *Eighteenth-Century Fiction* 24, no. 2: 161–70.

Burke, Tim. 2011. "Book Notes: Franco Moretti's *Graphs, Maps, and Trees*." In *Reading* Graphs, Maps, Trees*: Critical Responses to Franco Moretti*, edited by Jonathan Goodwin and John Holbo, 41–48. Anderson, South Carolina: Parlor Press.

Cain, William E. 1982. "The Institutionalization of the New Criticism." *Modern Language Notes* 97, no. 5 (December): 1100–20.

Clancy, Eileen. 2015a. "A Fabula of Syuzhet." *Storify* blog, https://storify.com/clancynewyork/contretemps-a-syuzhet, accessed 25 January 2016.

---. 2015b. "A Fabula of Syuzhet II." *Storify* blog, https://storify.com/clancynewyork/a-fabula-of-syuzhet-ii, accessed 25 January 2016.

Clement, Tanya. 2013. "Text Analysis, Data Mining, and Visualizations in Literary Scholarship." In *Literary Studies in the Digital Age: An Evolving Anthology*, edited by Kenneth M. Price and Ray Siemens. Modern Language Association Commons. http://dlsanthology.commons.mla.org/text-analysis-data-mining-and-visualizations-in-literary-scholarship/.

Darnton, Robert. 1982. "What is the History of Books?" *Daedalus*: 65–83.

DeWitt, Anne. 2015. "Advances in the Visualization of Data: The Network of Genre in the Victorian Periodical Press." *Victorian Periodicals Review* 48, no. 2 (Summer): 161–182.

Drucker, Johanna. 2009. "Entity to Event: From Literal, Mechanistic Materiality to Probabilistic Materiality." *Parallax* 15, no. 4. http://dx.doi.org/10/1080/13534640903208834.

---. 2011. "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly* 5, no. 1: http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

---. 2014. "Distributed and Conditional Documents: Conceptualizing Bibliographical Alterities." *MATLIT: Materialidades da Literatura/Materialities of Literature* 2, no. 1: 11–29.

Eggert, Paul. 2009. "The Book, Scholarly Editing and the Electronic Edition." In *Resourceful Reading: The New Empiricism, eResearch and Australian Literary Culture*, edited by Katherine Bode and Robert Dixon, 53–69. Sydney: Sydney University Press.

---. 2013. *Securing the Past: Conservation in Art, Architecture and Literature*. Cambridge: Cambridge University Press.

Elliot, Simon. 2002. "Very Necessary But Not Quite Sufficient: A Personal View of Quantitative Analysis in Book History." *Book History* 5: 283–93.

English, James. 2010. "Everywhere and Nowhere: The Sociology of Literature After 'the Sociology of Literature'." *New Literary History* 41, no. 2: v-xxiii.

Flanders, Julia, and Matthew Jockers. 2013. "A Matter of Scale." Faculty Publications – Department of English Paper 106, no. 24. http://digitalcommons.unl.edu/englishfacpubs/106.

Folsom, Ed. 2007. "Database as Genre: The Epic Transformation of Archives." *PMLA* 122, n. 5: 1571–79.

Freedman, Jonathan. 2007. "Whitman, Database, Information Culture." *PMLA* 122, no. 5: 1596–1602.

Frow, John. 2008. "Thinking the Novel: Review of *The Novel*, edited by Franco Moretti." *New Left Review* 49: 137–45.

Gibbs, Frederick W., and Daniel J. Cohen. 2011. "A Conversation with Data: Prospecting Victorian Words and Ideas." *Victorian Studies* 54, no. 1: 69–77.

Goldstone, Andrew. 2015. "Distant Reading: More Work to be Done." *Andrew Goldstone* blog, http://andrewgoldstone.com/blog/2015/08/08/distant/, accessed 25 January 2016.

Jockers, Matthew L. 2012. "500 Labeled Themes from a Corpus of 19th-Century Fiction." *Matthew L. Jockers* blog, http://www.matthewjockers.net/macroanalysisbook/macro-themes/, accessed 3 December 2014.

---. 2013a. *Macroanalysis: Digital Methods and Literary History*. Champaign: University of Illinois Press.

---. 2013b. "Confusion Matrices." *Matthew L. Jockers* blog, http://www.matthewjockers.net/macroanalysisbook/confusion-matrices/, accessed 3 December 2014.

Jockers, Matthew L., and David Mimno. 2013. "Significant Themes in 19th-Century Literature." *Poetics* 41: 750–69.

Leighton, Mary Elizabeth, and Lisa Surridge. 2009. "The Transatlantic *Moonstone*: A Study of the Illustrated Serial in *Harper's Weekly*." *Victorian Periodicals Review* 42, no. 3: 207–43.

Lohr, Steve. 2013. "Dickens, Austen and Twain, Through a Digital Lens." *New York Times*, January 26. http://www.nytimes.com/2013/01/27/technology/literary-history-seen-through-big-datas-lens.html?pagewanted=all&_r=0

Love, Heather. 2010. "Close but not Deep: Literary Ethics and the Descriptive Turn." *New Literary History* 41, no. 2: 371–91.

Liu, Alan. 2008. *Local Transcendence: Essays on Postmodern Historicism and the Database*. Chicago: University of Chicago Press.

Mak, Bonnie. 2014. "Archaeology of a Digitization." *Journal of the Association for Information Science and Technology* 65, no. 8: 1515–26.

McGann, Jerome. 1991. *The Textual Condition*. Princeton: Princeton University Press.

---. 2004. "A Note on the Current State of Humanities Scholarship." *Critical Inquiry* 30, no. 2 (Winter): 409–13.

---. 2005. "From Text to Work: Digital Tools and the Emergence of the Social Text." In *The Book as Artefact: Text and Border*, edited by Anne Hansen, Roger Lüdeke, Wolfgang Streit, Cristina Urchueguía, and Peter Shillingsburg, 49–62. Amsterdam: Rodopi.

---. 2007. “Database, Interface, and Archival Fever.” *PMLA* 122, no. 5 (October): 1588–92.

---. 2014. *A New Republic of Letters: Memory and Scholarship in the Age of Digital Reproduction*. Cambridge, Massachusetts: Harvard University Press.

McKenzie, D. F. 1999 [1986]. *Bibliography and the Sociology of Texts: The Panizzi Lectures*. London: British Library.

Moretti, Franco. 2000. “Conjectures on World Literature,” *New Left Review* 1: 54–68.

---. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London: Verso.

---. 2013a. *Distant Reading*. London: Verso.

---. 2013b. “‘Operationalizing’: or, the Function of Measurement in Modern Literary Theory.” *Stanford Literary Lab Pamphlet* 6 (December 2013): 1–13. https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf

---. 2013c. “The Bourgeois: Between History and Literature; Review and Interview by Karen Shook.” *Times Higher Education*, June 27, https://www.timeshighereducation.co.uk/books/the-bourgeois-between-history-and-literature-by-franco-moretti/2005020.article

Piepenbring, Dan. 2015. “Man in Hole: Turning Novels’ Plots into Data Points.” *Paris Review*, February 4, http://www.theparisreview.org/blog/2015/02/04/man-in-hole/.

Ross, Shawna. 2014. “In Praise of Overstating the Case: A Review of Franco Moretti, *Distant Reading* (London: Verso, 2013).” *Digital Humanities Quarterly* 8, no. 1. http://www.digitalhumanities.org/dhq/vol/8/1/000171/000171.html.

Rothman, Joshua. 2014. “An Attempt to Discover the Laws of Literature.” *New Yorker Magazine*, March 20. http://www.newyorker.com/books/page-turner/an-attempt-to-discover-the-laws-of-literature.

Schultz, Kathryn. 2011. “What is Distant Reading?” *New York Times Sunday Book Review*, June 24. http://www.nytimes.com/2011/06/26/books/review/the-mechanic-muse-what-is-distant-reading.html?_r=0.

Shillingsburg, Peter. 2010. “How Literary Works Exist: Implied, Represented and Interpreted.” In *Text and Genre in Reconstruction: Effects of Digitization on Ideas, Behaviours, Products and Institutions*, edited by Willard McCarty, 165–82. Cambridge: OpenBook Publishers.

Smith, David, Ryan Cordell, and Abby Mullen. 2015. “Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers.” *American Literary History* 27, no. 3 (August): online.

So, Richard, and Hoyt Long. 2013. “Network Analysis and the Sociology of Modernism.” *boundary 2* 40, no. 2 (Summer): 147–82.

St Clair, William. 2004. *The Reading Nation in the Romantic Period.* Cambridge: Cambridge University Press.

Suarez, Michael F. 2009. "Towards a Bibliometric Analysis of the Surviving Record, 1701–1800." In *The Cambridge History of the Book in Britain*, edited by Michael F. Suarez and Michael L. Turner, 37–65. Cambridge: Cambridge University Press.

Sunyer, John. 2013. "Big Data Meets the Bard." *Financial Times*, June 15. http://www.ft.com/cms/s/2/fb67c556-d36e-11e2-b3ff-00144feab7de.html

Trumpener, Katie. 2009. "Paratext and Genre System: A Response to Franco Moretti." *Critical Inquiry* 36, no. 1: 159–71.

Underwood, Ted. 2015. "A Dataset for Distant-reading Literature in English, 1700-1922." *The Stone and the Shell: Using Large Digital Libraries to Advance Literary History* blog, August 7. http://tedunderwood.com/2015/08/07/a-dataset-for-distant-reading-literature-in-english-1700-1922/.

Underwood, Ted, and Jordan Sellers. 2015. "How Quickly Do Literary Standards Change?" *figshare.* https://figshare.com/articles/How_Quickly_Do_Literary_Standards_Change_/1418394. Forthcoming in *Modern Language Quarterly* in 2016.

Wilkens, Matthew. 2013. "The Geographic Imagination of Civil War American Fiction." *American Literary History* 25, no. 4: 803–40.

NOTES

[1] While Jockers's claim to have identified six, maybe seven, plots in literature received significant public attention, the Syuzhet package he created for this analysis was subjected to substantial criticism within digital humanities, most notably from Anne Swafford, who questioned a number of its key assumptions and methods (see Clancy 2015a and 2015b for a summary of, and links to, relevant blogs in this debate).

[2] Ed Folsom's (2007) essay on database as a new genre offers another prominent example of this association of digital technologies with comprehensive and direct access to the literary historical record, while the responses to this essay highlight the inadequacy of such an approach (see especially McGann 2007 and Freedman 2007).

[3] Again demonstrating the difficulty Jockers (2013: 175) has in acknowledging the gap between comprehensive access and the access he attains, this claim that everything has been digitized is footnoted with: "No, not everything, but compared to 1988, yes, everything imaginable."

[4] In contrast to the other co-authors, no specific role or insight is ascribed to Moretti in this paper.

[5] The one exception to this failure to publish underlying data comes in a pamphlet they author collaboratively with others (see Allison et al. 2014).

[6] Elsewhere Jockers confirms the hint he gives in *Macroanalysis* that he does not publish data because it is derived from proprietary collections (Jockers and Mimno 2013: 752); but this does not explain why he cannot name the authors and titles studied or provide textual data at the level of word frequencies.

[7] In explaining the scope and intention of "macroanalysis," Jockers (2013a: 28, 9) cites this passage, as well as Russian formalist Juri Tynjanov's assertion that "one cannot study literary phenomena outside of their interrelationships."
[8] Book history is premised on a systemic conception of print culture, as manifested, probably most famously, in Robert Darnton's (1982) "communication circuit."
[9] Bode 2012. Wilkens 2013 also demonstrates this approach.
[10] For foundational work in this area see McKenzie 1999 [1986] and McGann 1991.
[11] For example, Drucker 2009 and Eggert 2013.
[12] The technique misclassifies 33 percent of works in terms of nationality; 14 percent of works in terms of gender; and an unspecified proportion of works in terms of chronology.
[13] Again indicating Moretti's disinterest in the particulars of his datasets, he claims the graph depicts an increase from 5 to 50 novels per year in all national contexts, whereas the data he displays has British novels increasing to 30 titles, Japanese and Spanish novels to a little over 40, Italian novels to 35, and Nigerian novels to only 25 titles per year (Moretti 2005: 6).
[14] Demonstrating the spiraling (though in this case, arrested) abstraction liable to occur when the rhetoric of text is applied to a large number of literary works, in a published debate Jockers (Flanders and Jockers 2013) remarked that he had decided to cease collecting "texts" when he had 4,700, but upon reaching that number and looking in more detail at the records, realized he had only 3,346 because: "the materials my colleagues and I had collected included many multi-volume novels that had not been stitched together and also a good number of duplicates that we had acquired from different sources."
[15] For an influential discussion of the interpretive status of data see Drucker 2011.
[16] Longstanding projects in this vein include the Rossetti Archive (http://www.rossettiarchive.org/) and the Orlando Project (http://orlando.cambridge.org/), with digitization of the full-run of the *Western Home Monthly* a more recent, and ongoing, example (http://modmag.ca/whm/).
[17] This approach characterizes the essays in the first half of *Virtual Victorians* (Alfano and Stauffer 2015).
[18] Neither So and Long (2013) nor DeWitt (2015) discuss the gaps in the digital collections they use or in the datasets they construct on that basis; Smith, Cordell, and Mullen (2015) note the inevitable difference between the newspapers published and the mass-digitized collection they use to investigate reprinted texts, but do not characterize the nature and extent of these gaps.
[19] More recently, McGann (2014: 26) described the scholarly edition as "a model, a theoretical instantiation, of the vast and distributed … network in which we have come to embody our knowledge."
[20] Mak (2014) provides an excellent model for the first three stages of this history of transmission, in her "archaeology" of Early English Books Online.
[21] For fuller discussion of these issues see Bode and Hetherington 2014.
[22] Underwood, "Dataset."