Preprint – Forthcoming in *Modern Language Quarterly* 78.1 (2017)

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

Katherine Bode

In a 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 L. 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 few book-length contributions to the field (Jockers 2013a; Moretti 2005, 2013a), but their work is reported on in major public forums such as the *Paris Review*, the *Financial Times,* the *New York Times*, and the *New Yorker* (Lohr, 2013; Piepenbring 2015; Rothman, 2014; Sunyer, 2013; Schultz, 2011), 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, especially, has also received its share of criticism, following several lines. The most resistant scholars maintain that data are inimical to literature, which only close reading can explore in its nuance and complexity (e.g. Trumpener 2009). James English (2010: xii, xiii) attributes this response to the discipline's foundationally "negative relationship" with "counting," and to Moretti's role in widening 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 quantification inevitably abstracts and simplifies complex phenomena (e.g. Love 2010: 374). Regarding Moretti's tendency to "overestimat[e] the scientific objectivity of his analyses" (Ross 2014), some perceive his 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 "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 is inherited from the New Criticism: contrary to the prevailing view, close reading 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, I do not focus on the individual literary works that 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 the core practice in Moretti's and Jockers's work – modeling literary systems – and to the mass-digitized collections that make this historical formation amenable to analysis.

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 for literary history, and was unrelated to either data or computation (Moretti 2000). In *Graphs, Maps, Trees* (Moretti 2005), literary history was foregrounded, and the "units much smaller or larger than the text," theorized in "Conjectures on World Literature" (Moretti 2000: 57), were 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" (Margaret 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 (2014) 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 a 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. In the first chapter of *Graphs* Moretti (2005: 3, 9, 30) repeated referred to literary data as "facts," "ideally independent of interpretations"; as "*data*, not interpretation"; and as "useful because they are independent of interpretation." On this basis he 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 a historical explanation to the results, underpin Moretti's contention to explore "the literary field as a whole" (67).

Missing from extant critiques is Moretti's neglect of the scholarly infrastructure supporting 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 the latter study – showing the number of British novels – stops in 1836 (where the other graphs extend to 1850), and a footnote comments, "it seems very likely that Andrew Block's bibliography significantly overstates the number of novels published after that date" (2013a: 188). Yet acknowledging that a particular set of literary data is the outcome of a ("significantly" flawed) "interpretive encounter" affects neither Moretti's rhetoric nor his subsequent analysis: the chapter still claims to "read the entire volume of the literary past" (58), and while the data are absent from figure 7, Block's bibliography is the only source for titles published from 1836 to 1850. Moretti proceeds, in other words, to analyze 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 such collections the rhetorical, if not the primary analytical, focus of *Distant Reading,* Moretti (2013a 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." Moretti (2013c) reinforces his manifest disregard for the specifics of disciplinary infrastructure by aligning 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 interpretive 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 fueled 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," "big data" is supposedly separate from human involvement and thus offers "comprehensive and definitive" historical facts (31). According to Jockers, 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" (7–8). As Jockers says of one of Moretti's analyses, this unprecedented and uninterrupted access to the matter of literary history "leaves little room for debate" (20). 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 … beneath the mass of gravel layered above." Working with the "gravel" of literary history enables Jockers "to unearth, for the first time, what these corpora really contain," a metaphor that conflates analysis with achieving complete access.

Where Moretti occasionally acknowledges limitations in his data (before proceeding with analyses regardless), Jockers (28) maintains that any "leap from the specific to the general" is flawed. Only in the book's final chapter does he admit the obvious gap between his datasets and the "population" of nineteenth-century novels, describing his largest "corpus of 3,346 books" as "incomplete, interrupted, haphazard" (172). This concession that "the comprehensive work is still to be done" (172) 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" (171) – literary scholars are advantaged over evolutionary biologists "in terms of the availability of our source material" (172). 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 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" (175).[3]

The most recent collaboration from that group departs, which Jockers is no longer part of, 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).[4] Although imagining that the "convergence of the 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. Yet 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 to generate "a random sample" of what was published, the authors overlook the fact that both digitized collections and analog bibliographies 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 revealing its datasets.[5]

Moretti often references his sources of data – chapter one of *Graphs*, for instance, begins by listing the bibliographies it draws on (Moretti 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." Moretti, however, does not share his data. Jockers occasionally publishes the results of data analysis – such as the 500 themes developed from topic modeling and presented as word clouds on his website (Jockers 2012) – but he does not provide the textual data analyzed, even at the level of word frequencies,[6] and he is significantly less open than Moretti about the sources 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 3,346 (identified in the context of reporting confusion matrices) (Jockers 2013b).

In Moretti's case, one might suppose it possible to reconstruct the datasets from cited sources. But his account (in an appendix to *Graphs*) of creating the dataset for "British novelistic genres, 1740-1900" highlights why it is not feasible. There he describes his periodization as "not always explicit" in the bibliographies (Moretti 2005: 31), 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 particular texts of literary works, but highly prepared – or pre-interpreted –

selections from those texts). Far from an incidental oversight, the refusal to publish datasets maintains the fiction that literary data are prior to interpretation: it removes the need either to describe the procedures for collecting, cleaning, and curating datasets, or to expose the inevitably selective and limited collections resulting 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 by the selections and amplifications that that framework produces. Accordingly, two scholars can read the same dataset, like the same literary work, and derive different meanings from it. 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, transcribing them, analyzing those transcriptions, publishing the findings, and asserting that they demonstrate a "definitive" new perspective on the literary field, without enabling anyone else to read the transcriptions (or, in Jockers's case, without revealing the titles of most of the original documents).

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 primary units 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, Moretti and Jockers ultimately fail to capture the historical nature of literary works and how they connect to produce literary systems.

Moretti and Jockers typically define literary works by 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 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 of 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 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 [Bode 2012]),[9] of using publication data to test existing perspectives on literary history. 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," Jockers (2013a: 36) challenges the notion of a "lost generation" of Irish-American authors 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 existing arguments, and is highly innovative in this respect. His study of titles, for instance, investigates a category of literary data that had not, as far as I know, been subject to synoptic, stylistic analysis (Moretti 2013a: 179–210); more broadly, he combines multiple bibliographies to explore relationships between different literary-historical 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. In *The Reading Nation in the Romantic Period* William St Clair aptly diagnoses the limitations of this approach. Like Moretti and Jockers, St Clair (2004: 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, frequently include "texts written or compiled long ago and far away" (3), and some works are inevitably 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 A. 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 disregarding the varied profiles and presences of literary works in history. The popularity of different literary works when they are published and in subsequent generations accords them "a massively different footprint" in history, altering their influence and therefore their meaning. But commercial success is not the only relevant factor. As demonstrated 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 (e.g., Drucker 2009; Eggert 2013), Moretti and Jockers construct literary systems as composed 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 rely on basic features of new literary production to constitute both the problem and its explanation. *Macroanalysis* purports to investigate "the context in which [literary] change occurs" (Jockers 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 [134]) 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,[11] 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 few predetermined categories, Jockers is confined to stating their presence. He cannot offer any alternative influences, nor can he 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 his various chapters: "The linguistic choices an author makes are, in some notable ways, dependent upon, or entailed by, their genre choices" (Jockers 2013a: 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 Jockers 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 "are both central to his argument and absent from his evidence" (DeWitt 2015: 162). 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 figure 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,[12] 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 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 way that 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).

Yet the same problem – of assuming the shape of the past from that of the present – occurs on 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 (2013a: 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) who 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 his 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 to 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 (which are reliant on identifying attributes that can be represented in uniform fields), reduction and abstraction characterize all analysis (even so-called close readings interpret not 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.

Where distant reading and macroanalysis are celebrated – or decried – for their departure from "close reading," these approaches share a disregard for textual scholarship, and an 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, or 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 to 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 subordinated the historical – biographical, material, sociological – concerns of previous scholarship to the text itself. 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 that 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) that are 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, composed 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, and in a manner rendered only more abstract by the large number of "texts" under consideration.[13]

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 that is contained in the material form assessed by the critic. Yet even with these protections, the text's centrality and assumed singularity – and the dissociation from the literary work's complex historical existence that this conception produces – can diminish the capacity of close reading to investigate literary history. For example, in discussing Wilkie Collins's *Moonstone*, Mary Elizabeth Leighton and Lisa Surridge (2009: 207) note the varied interpretations of its meaning for nineteenth-century readers: from tale of "imperialist panic" to class critique. While all critics believe that "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 a 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 – by Moretti himself and Jockers, among others – to integrate "traditional and computational approaches" (Gibbs and Cohen, 2011: 70; Jockers 2013a: 26; Moretti 2013a: 181). Understood in terms of the different perspectives that the two approaches 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, more specifically the fact that producing such an object is itself a critical and interpretive enterprise.[14] 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 proved so difficult, in practice if not in theory, to integrate data-rich and traditional 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,[15] or using mass-digitized collections to historicize particular titles and authors,[16] a number of projects consider 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; DeWitt (2015) studies genre formation in terms of 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 (Cordell 2015; Smith et al. 2015). Based on the success of their machine learning model in distinguishing reviewed from unreviewed poetry titles Ted Underwood and Jordan Sellers (2016) challenge the longstanding view that standards governing literary prestige changed suddenly at the end of the nineteenth century.

All 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 by publishing them. Discussing their work on literary prestige, Underwood and 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. Underwood and Sellers publish the datasets and code underpinning their work; in collaboration with HathiTrust, Underwood (2015) has also published word counts by subgenre, with yearly summaries.

While the departure from Moretti's and Jockers's approach is significant, these other projects have 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 Stanford Literary Lab pamphlet *Canon/Archive* 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.[17] Underwood and Sellers recognize the difference between the historical context investigated and the mass-digitized collection used in that process, and they 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 (Underwood and Sellers 2016: 338).

Yet even this project is equivocal in characterizing the broader relationship between historical literary systems and our disciplinary infrastructure. 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 record nor the digital one offers 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 a 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 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 to 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 extending the insights of such arguments to the discipline of literary history as a whole.

In making this claim, I am not proposing that scholarly editions of literary works and of literary systems 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 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 of 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 (2014: 117) words, a "hypothetical platform" for historical inquiry, one that both demonstrates and provides a pathway through the instability of the "textual condition".[18] This capacity inheres in the interrelationship between the edition's parts: the historical introduction, the critical apparatus, and the curated text. The first two describe and justify the editor's engagement with the documentary record constitutive of the literary work, specifically with the inevitable gaps and uncertainties that that engagement exposes. The latter represents the outcome of the extended critical encounter, offering to those who accept its tenets – or who lack the expertise to engage with the literary-historical record in this manner – a basis of analysis.

For the scholarly edition of a literary system, the historical introduction and the critical apparatus describe 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 projects in data-rich literary history – such an introduction 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;[19] and the additional transformations wrought by the editor's extraction, construction, and analysis of those data. Where the historical introduction explains and justifies the guiding principles and general composition of the modeled literary system, the critical apparatus details particular decisions and arguments made in data construction.

The curated dataset manifests – demonstrates and, more particularly, publishes – the outcome of this careful exploration of the successive acts of production and reception 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 a historical literary system for others to investigate, for either traditional literary-historical or data-rich research. Where "the stylistic protocols of literary criticism" mean that issues deemed methodological are relegated to footnotes or "methodological caveats" (Underwood and Sellers 2016) – as if they qualified rather than constituted the basis of the arguments offered – a scholarly edition of a literary system provides a dedicated format for demonstrating and justifying the foundational argument of data-rich literary history: the modeling of a literary system.

In its stability, the curated dataset for a scholarly edition of a literary system resembles the word counts that Underwood has released in conjunction with HathiTrust. Such stability is vital: where all documents and collections are "transactional" entities (Drucker 2014: 12) – 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 collection and the curated dataset 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 curated dataset of a scholarly edition of a literary system differs from the word counts offered by the Underwood/HathiTrust, not only in its detailed historical

and critical accompaniments, but in two other, important ways. First, and most basically, where the latter is usable only by those with programming expertise, or access to it, a curated dataset is accessible to all literary scholars through 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" (Underwood 2015), the curated dataset of a scholarly edition already embodies an argument regarding historical relationships between literary works. Clearly, such a dataset 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 form of literary works: predominantly English-language American editions). The curated dataset of a scholarly edition of a literary system is intended not as raw data. Rather, with its associated historical introduction and critical apparatus it constitutes a historically contextualized model of literary historical events and connections, and an interpretive intermediary between increasingly complex, digital disciplinary infrastructure and the requirements of literary-historical analysis.

For the scholarly edition of a literary system I am building – of fiction published in nineteenth-century Australian newspapers – my historical introduction traces 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 that my data model presents.[20] However, as in a scholarly edition of a literary work, the issues foregrounded by such an introduction 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.

My curated dataset, when published, will present detailed bibliographical metadata and digitized text for approximately sixteen thousand stories published in nineteenth-century Australian newspapers. Many of these stories are not known to have appeared in Australia; some of them are not known to have been published at all. I will publish these 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 composition of a curated dataset will reflect what the editor (or editorial team) perceives as most relevant to understanding a particular literary system. 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 and information relevant to the publication event (e.g., changes in the title or the various designations of authorship). The critical apparatus is published alongside the curated dataset, and explains and justifies the basis of the bibliographical and textual data presented (e.g., supplying information sources for authorship, references for textual derivations, or additional sites of publication for the fiction discovered).

For some researchers, a scholarly edition of a literary system will function much 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 new works, who knows how many, that were not previously known to literary historians but were 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, a scholarly edition of a literary system will provide a carefully, consistently, and – by the historical introduction and the critical apparatus – explicitly historicized digital collection for this task. For researchers interested in analyzing large-scale trends in the publication, circulation, and reception of literary works, such an edition will provide a rigorously constructed and explained "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 it can attend to various features and dimensions of the modeled literary system. Alan Liu's (2008) notion of "contingency" encapsulates the resulting analytic mode. Describing the relationship of postmodernism historicism to the database, Liu notes that, in neither form does one ask: what is the meaning of this whole? Rather, the question

becomes: 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 a 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 am advocating 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 a 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 historical introduction, critical apparatus, and curated dataset, a scholarly edition of a literary system offers an explicitly theorized and practical framework for exploring the multiple instances of production and reception that are constitutive of the data-rich models analyzed, and for assessing and managing the inevitable contingency of that situation for historical analysis.

As the basis of 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 that is foundational to the New Criticism without benefiting from the institutional and infrastructural protections afforded 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.

References

Alfano, Veronica, and Andrew Stauffer, eds. 2015. *Virtual Victorians: Networks, Connections, Technologies.* Houndmills 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*. January. Stanford Literary Lab Pamphlet Series. http://litlab.stanford.edu/LiteraryLab Pamphlet11.

Allison, Sarah, Ryan Heuser, Matthew Jockers, Franco Moretti, and Michael Witmore. 2011. *Quantitative Formalism: An Experiment*. January. Stanford Literary Lab Pamphlet Series.. http://litlab.stanford.edu/LiteraryLabPamphlet1.

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.

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

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/.

Cordell, Ryan. 2015. "Reprinting, Circulation, and the Network Author in Antebellum Newspapers." *American Literary History* 27, no. 3: 1–29.

Darnton, Robert. 1982. “What is the History of Books?” *Daedalus* 111, no. 3: 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. DOI: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.” *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.

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

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/.

Jockers, Matthew L. 2012. “Five Hundred Labeled Themes from a Corpus of Nineteenth-Century Fiction.” Matthew L. Jockers (blog). http://www.matthewjockers.net/macroanalysisbook/macro-themes/.

---. 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/.

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

Jockers, Matthew L., and David Mimno. 2013. "Significant Themes in Nineteenth-Century Literature." *Poetics* 41, no. 6: 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*. December. Stanford Literary Lab Pamphlet Series. https://litlab.stanford.edu/LiteraryLabPamphlet6.

---. 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*, 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.

Smith, David, Ryan Cordell, and Abby Mullen. 2015. “Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers.” *American Literary History* 27, no. 3 (August). alh.oxfordjournals.org/content/27/3/E1.full.

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.

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 (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. May 19. DOI: 10.6084/m9.figshare.1418394.

---. 2016. “The Longue Durée of Literary Prestige.” *Modern Language Quarterly* 77, no. 3: 321–44.

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, founder and director of DASH Lab (Digital Arts, Sciences, and Humanities Lab) at SUNY New Paltz, who questioned a number of its key assumptions and methods (see Clancy 2015a, 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 his essay highlight the inadequacy of such an approach (see esp. Freedman 2007 and McGann 2007).

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

[4] No specific role or insight is ascribed to Moretti, unlike the other coauthors, in pamphlet 11.

[5] The one exception to this failure to publish underlying data is a pamphlet on which Moretti and Jockers collaborated with others (see Allison et al. 2011).

[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) quotes this passage, as well as the 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: 67) "communication circuit."

[9] Wilkens 2013 also demonstrates this approach.

[10] For foundational work in this area see McKenzie 1999 [1986] and McGann 1991.

[11] The technique misclassifies 33 percent of works in terms of nationality; 14 percent of works in terms of gender, and an unspecified percentage of works in terms of chronology.

[12] Again indicating Moretti's neglect of the particulars of his own datasets, Moretti claims the graph depicts an increase from 5 to 50 novels per year in all national contexts, whereas the

data he displays have British novels increasing to thirty titles, Japanese and Spanish novels to few over forty, Italian novels to 35, and Nigerian novels to only twenty-five (Moret,i 2005: 6).

[13] 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 remarked in a published debate that he had decided to cease collecting "texts" when he had 4,700 but, on reaching that number and looking in more detail at the records, realized that 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" (Jockers and Flanders 2013: 24).

[14] For an influential discussion of the interpretive status of data, see Drucker 2011.

[15] 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* an ongoing example (http://modmag.ca/whm/).

[16] This approach characterizes the essays in the first half of *Virtual Victorians* (Alfano and Stauffer 2015).

[17] Neither So and Long (2013) nor DeWitt (2015) discusses 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 they do not characterize the nature and extent of these gaps.

[18] McGann (2014: 26) also describes the scholarly edition as "a model, a theoretical instantiation, of the vast and distributed … network in which we have come to embody our knowledge."

[19] In her "archaeology" of Early English Books Online, Bonnie Mak (2014) provides an excellent model for the first three stages of this history of transmission.

[20] For fuller discussion of these issues, see Bode and Hetherington 2014.