A World of Fiction

Bode, Katherine

Published by University of Michigan Press

Bode, Katherine.
In making the case for and describing a new object of analysis for data-rich literary history, based on the scholarly edition, this chapter builds on McGann’s call for “philology in a new key,” most recently in *A New Republic of Letters*, and on Eggert’s work on the scholarly edition as a model for conservation broadly, in *Securing the Past*. Like McGann, I propose that literary scholars respond to the challenges of digitization by employing the theoretical and practical framework through which they have long negotiated the complexity of the documentary record. I also take seriously McGann’s insistence on the need to develop digital environments that demonstrate the histories of production and reception, including the current moment of interpretation, for the documentary records they model. Like Eggert, I apply this framework beyond the individual literary works that McGann maintains are the essential basis of an “object-oriented and media approach to the study of literature and culture” (*New* 3). In using the scholarly edition as a basis for modeling literary systems and for investigating the mass-digitized collections that make such historical formations amenable to analysis, I also seek to chart a path beyond the polemics of both “distant reading” and “macroanalysis” on the one hand and McGann’s...
“new philology” on the other, and thus beyond the view of digital 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).

Multiple projects already model historical literary systems with attention to relationships of production and reception and to the meanings these connections signify. Explicit engagement with modeling, as the method is theorized in digital humanities, is essential for data-rich literary history, but modeling alone is insufficient. Modeling recognizes data as constructed, but only by the individual scholar; it does not provide a mechanism to interrogate the history of transmission preceding and perpetuated by the scholar’s engagement with the documentary record, including in its mass-digitized forms. The framework of the scholarly edition meets that challenge, presenting a structure to negotiate the incomplete and transactional nature of the documentary record and to represent the outcomes of that process. Adapted to the literary system, it offers a reliable foundation for data-rich literary history and for extending the insights gained from that field’s engagement with emerging digital infrastructure to the broader discipline.

Although Moretti’s and Jockers’s work is often taken as definitive of data-rich literary history (even of digital literary studies), many projects model literary systems in significantly more nuanced ways. Far from assuming that literary works published around the same time and by authors of the same nation automatically constitute a system, these projects investigate interconnections—temporal, spatial, and social—between literary works in the past. To put this practice in the terms of textual scholarship, they depart from the ideal or regulatory category of the literary work to investigate the ways and forms in which literature has circulated and generated meaning. Where Moretti and Jockers assume a relationship between what was published and what was read, these projects consider forms and relationships of production and reception. As with this book, literary data derived from periodicals is often used—and such research will be my focus in this chapter. Because periodicals are published for particular readerships at particu-
lar times, such data have clear temporal, spatial, and social dimensions; periodical collections are also increasingly and extensively digitized, making them amenable to data-rich scholarship. However, literary systems could be modeled in historically meaningful ways using any digital collection(s) or other disciplinary infrastructure, as long as a schema for production and reception is applied to sample selection and analysis.

Some data-rich literary history projects use periodical data to model literary systems based on the authors and titles presented to particular readerships. Richard So and Hoyt Long employ analog bibliographies of poetry in American and Japanese periodicals—and construct their own bibliography for the Chinese context—to identify authors published together. On this basis they propose and contrast “the collaborative networks that underwrote the evolution of modernist poetry” in these countries (148). As So and Long note, because poetry periodicals served “as the institutional sites through which an expanding market for avant-garde poetry was stratified and differentiated along aesthetic, ideological, racial and even geographical lines of affiliation,” data derived from them have an “inherently social dimension” that supports exploration of the historical and structural conditions under which modernist poetry emerged (158). The historical newspapers and magazines analyzed in the Viral Texts project, led by Ryan Cordell and David Smith, provide similar contextual information. Employing an algorithm to identify republished passages, the Viral Texts project discovers works in a range of genres—including many not traditionally considered by literary historians (jokes, recipes, “listicles”)—where reprinting indexes popularity as well as personal and structural connections between newspaper editors (Cordell, “Reprinting”).

Other projects model literary systems using more direct evidence of reception. Ed Finn explores the “public literary action” and resulting positions in the American literary field of four prominent authors—Thomas Pynchon, Toni Morrison, David Foster, and Junot Díaz—based on the authors and titles they are linked to in reviews in key American periodicals, as well as in book recommendations and reviews on Amazon. Although Internet communities are “irrepressibly international,” as Finn notes, such data signifies how contemporary readers, mainly though not exclusively in the United States, position these authors and their works (48). DeWitt uses a similar approach to explore genre
formation in the nineteenth century. Noting her methodological debt to Finn, DeWitt begins with seven prominent theological titles and conducts keyword searches in six mass-digitized collections to gather colocation data: titles mentioned in the same book reviews in Victorian periodicals. The resulting analysis characterizes genre as “a classification created retrospectively by readers to make sense of a field of novels” (177). Although DeWitt’s model has social, temporal, and spatial parameters, she notes the potential for more detailed investigation with additional data that “differentiate between different kinds of articles, taking into account the date or location of publication as well as the sorts of periodicals in which they appear” (176).

Ted Underwood and Jordan Sellers’s study of conceptions of literary prestige for poetry published between 1820 and 1919 is also based on periodical reviews: namely, 360 poetry titles reviewed in prominent American and British periodicals (“Longue”). But rather than investigating their contents, Underwood and Sellers use these reviews to identify a sample of reviewed poetic works. They then compare word usage in those works to that in 360 nonreviewed titles, selected randomly from the HathiTrust Digital Library. While this project—like DeWitt’s—arguably conflates cultures of reviewing and constructions of literary value in two different contexts—America and Britain—its focus on titles discussed at particular times, and reviews as a marker of cultural distinction, imbues the data with clear temporal and social dimensions.

These projects recognize that literary works do not exist in a single time and place but accrue meaning in the multiple contexts in which they are produced and received. And they create models based on that understanding to investigate key issues in literary history relating to production, circulation, and reception, ranging from the nature and development of literary movements and genres to the formation of authorial identities and cultural value. In some cases they overturn core assumptions in literary history. Thus the capacity of Underwood and Sellers’s machine-learning model to discern reviewed from unreviewed poetry titles enables them to demonstrate that standards governing literary prestige developed gradually over the nineteenth century, rather than changing suddenly at the end of it (“Longue”).

If sometimes implicitly, these projects highlight the importance for data-rich literary history of the computational method of modeling. Described most fully for the digital humanities in Willard McCarty’s
*Humanities Computing*, modeling is not simply a means of representing cultural artifacts or concepts in a form amenable to computational analysis. By foregrounding the manipulation and refinement of data, modeling produces an “experimental device” (27) or “pragmatic instrument of investigation” that supports an ongoing “process of coming to know” (36). This method enhances knowledge because these acts of representation and refinement expose the researcher’s assumptions about the cultural artifact or concept under investigation and enable those assumptions to be clarified or challenged.

One contribution this book aims to make is to expand and enrich the application of modeling for data-rich literary history by connecting it explicitly to descriptive bibliography. Moving beyond basic enumerative information, this approach investigates relationships of production and reception by describing, manipulating, and refining—that is, by modeling—details of the documentary forms and historical relationships of literary works in the past. Chapter 3 presents the outcome of this practice, identifying the multiple fields I use to describe fiction in nineteenth-century newspapers. Here, two instances of features modeled—and their relationship to production and reception—will help to indicate the value of this practice for data-rich literary history.

The title changes that fiction underwent when published and republished in Australian newspapers offers a relatively straightforward example of this integration of modeling and descriptive bibliography. Collecting all manifestations of a literary work under the same title, though standard in enumerative bibliography, obscures how this documentary feature was constituted by different newspaper editors and syndicators and experienced by different reading communities. Modeling fiction with respect to the various titles used and the common title uniting related manifestations of a work supports insights into publication and reception: for instance, how editors altered the works they published and what this implies about their understandings of the readerships they served. It also enables analysis of the systems through which fiction circulated, in that collating multiple versions of a title indicates the works that were widely published in the Australian colonies while offering clues as to the mechanisms of that circulation: for instance, variations in a title preclude the use of certain types of (ready printed and identical) newspaper supplements as the means of distributing fiction.
Authorship was one of the most complicated features of this literary system to model. Not only were many titles published anonymously but, as chapter 4 explores in depth, some were published with attribution in one instance, and without attribution in another; some anonymously published authors remain unknown today, while others have been identified; and in many cases, the paratext of stories contained information about authorship, regardless of whether the work was published anonymously or with a named author. I embedded this information in the model through such fields as “publication author” and “author,” to differentiate the form of attribution used in the newspaper from a single name uniting an oeuvre; “signature,” to identify other titles described as “by the author of” the story within the publication event; and “inscribed gender” and “inscribed nationality,” to depict information in subtitles and other aspects of the publication event designating the story as by a man or woman and from a particular place. Detailed modeling of authorship enabled a significantly more nuanced representation and investigation of this literary-historical phenomenon than would be possible with a single author name. It also supported analysis of reception. By indicating features of authorship highlighted by different newspaper editors and syndicators, for example, it suggests ideas about cultural value informing literary culture in the colonies.

Like other forms of textual scholarship, bibliography is often perceived as precritical and procedural. But its descriptive form has always approached bibliographical data as “capta, taken not given, constructed as an interpretation of the phenomenal world, not inherent in it” (Drucker, “Humanities” np). Modeling is likewise based on an understanding of data as contingent and constructed. Equally important, the two practices are underpinned by the same dual imperative: to enact and to enable critical analysis. Integrating modeling and descriptive bibliography—or more specifically, using descriptive bibliography as a framework for modeling and modeling as a method for extending bibliographical knowledge—can support detailed and nuanced representations of literary systems that explore the existence of literary works in the past and support future investigations of those works and systems. While data-based approaches to literary history are often accused of simplifying understandings of the past, this approach offers not only nuance and complexity but a means of refiguring and refining the terms in which literary history is pursued.
But however essential it is for data-rich literary history, modeling cannot be the sole foundation for the field. Models of literary systems are not simply arguments about the existence of and connections between literary works in the past; they are arguments made with reference to the disciplinary infrastructure—the bibliographies and collections, analog and digital—that transmit evidence of past works and relationships to the present. Modeling, even when integrated with descriptive bibliography as I have described, does not reflect on this transmission. Models embody a scholar’s arguments, whereas disciplinary infrastructure is an effect of multiple arguments: a sequence of assumptions, decisions, representations, and remediations. Such histories of transmission shape how the researcher can explore, and what she can know of, the historical context that disciplinary infrastructure appears to represent. To adequately perform literary history, data-rich projects must investigate these histories of transmission and how they constitute the documentary record.

II

The main issues foregrounded in discussions of digital collections, particularly mass-digitized ones, are scale and access, for obvious reasons. Digital collections enable new forms of access to documentary records through novel organizations of bibliographical data and full-text searching. Often, digital collections bring a larger number of related documents together than is held in any single analog collection. Yet inevitably, digital collections of historical documents are partial representations of partial records. While many data-rich literary history projects are moving toward more critical assessments of the disciplinary infrastructure they work with, the complexity of such infrastructure and its implications for research practices are yet to be fully appreciated and answered.

The distinctions or gaps between the context signified by collections and the exemplars used in signification might partly arise from, but are not simply the consequence of, successive exclusions of documents, as the Stanford Literary Lab Pamphlet 11 suggests. In chapter 1 I noted that, in defining “the published,” “the archive,” and “the corpus” as progressively smaller selections, those authors admit the constructed nature
of literary data. Yet they also argue that mass-digitization largely avoids those exclusions, such that “the corpus of a project can now easily be (almost) as large as the archive, while the archive is itself becoming—at least for modern times (almost) as large as all of published literature” (Algee-Hewitt et al., 2).

Even with their account of the considerable practical challenges involved in accessing versions of specific literary works, this description of mass digitization drastically diminishes the degree of exclusion involved in constructing such collections. Digitization only captures certain features of documentary texts, and the documents digitized are only some of those that existed and some of those that were collected by an institution or group of institutions. For mass-digitized periodical collections, for instance, the documents digitized are not only those that were collected, at set times and for specific reasons, but usually those that were also microfilmed for access and preservation. Digitized collections are partial in another way, in that combining the holdings of multiple analog collections tends to obscure the individual histories of the contributing collections and their implications for the form, scope, and critical capacity of the resulting digital one.

_Pamphlet 11_ belies further distinctions between published, collected, and analyzed documents created by their representation and remediation: that is, by the successive acts of production and reception, critical and technical, that produce digital collections. Collections have always been constituted in this way. In analog collections, documents are represented and remediated through the cataloging systems that organize holdings and the interfaces that interpret them: the card catalogs, special collection indexes, or online library catalogs that provide a method of searching and a type, form, and detail of metadata. As Eggert notes, a library or archive, “far from acting as a neutral frame” for the documents it transmits, “wraps them up in a relationship to the intended viewer” that shapes what is discovered and how it is understood (Securing 13–14). A key difference between analog collections and digital ones is that literary historians rarely, if ever, treat the former (a given library, for instance) as proxies for literature as it circulated and was understood in the past, whereas digital collections such as Google Books or HathiTrust are sometimes assumed to be representative in this way.

This perception persists even as the construction of digital collec-
tions constitutes a radical increase over analog ones in the degree and complexity of the representations and remediations involved. In addition to the history/ies of transmission of the analog collection(s) from which they are created, digital collections are the effects of multiple, intersecting structures and systems. Chief among an extensive array of these are the selection criteria employed in defining the digital collection and the technical workflow involved in implementing them; the metadata standards used to describe documents and how these are incorporated into the collection interface; the quality of (microfilming and) digitization and capacity of optical character recognition (OCR) to produce searchable text; the nature of the search algorithm(s) used to order results; and the interface(s) available for accessing the collection and the type(s) of access these support. The “corpus” selected from this collection and used for analysis is, in turn, shaped by the type and form of data that can be exported. All of these factors have their own constitutive effects, and all of them occur before the scholar begins the modeling process described in the previous section. Digital collections are also more mutable than analog ones, meaning that the results of a search conducted today might be very different from the same search conducted tomorrow.

Due to their multiplicity and complex interaction, the components involved in producing digital collections expand access to the historical record in certain ways, even as they increase the likelihood of unrealized and significant disjunctions between the access we intend and the access we achieve. Digital humanities scholars recognize that digital infrastructure shapes knowledge production, and digital literary historians have responded with explicitly curatorial approaches to constructing and exploring digital documents and collections. This approach is, for instance, definitive of the Text Encoding Initiative—an international consortium developing standards for the representation of texts in digital form—and of digital scholarly editing broadly.

A curatorial approach also defines projects that construct digital collections to represent particular literary works, authors, and/or publications in their historical context, including the Rossetti Archive, the Orlando Project, and the digitization of the Western Home Monthly. Such projects attend to the specificity of collections from which documents were gathered, the different documentary manifestations incorporated, and the partiality of the digitized record, and they devise methods
for representing and managing these features. These same issues are foregrounded in a growing body of research that uses these digital collections and others to historicize individual titles or authors. Recent work in this area advocates strategies for identifying and managing gaps in digital collections, including visualizing absences (Klein) and moving back and forth between digitized and analog collections (C. Robinson). In an approach similar to the one I am proposing, though focusing on particular authors and works rather than literary systems, other projects construct dedicated, scholar-built resources to circumvent the search algorithms organizing engagement with mass-digitized collections and/or to support forms of experimental and speculative analysis of literary data (Brown; Sinatra).

These digital humanities projects highlight four features that I believe should also underpin the modeling of literary systems in data-rich literary history. First is a critical assessment of the relationship between the historical context analyzed and the digital collection(s) used for analysis; second is detailed attention to the relationship between the documents included in the digital collection(s) and the terms in which they are represented; third is explicit discussion of the means by which data are extracted and modeled; and fourth is a published record of data arising from that extensive history of transmission. Existing projects in data-rich literary history often (though by no means universally) demonstrate the second and third of these features. The need for data publication, and for platforms and modes of review to support it, is also increasingly recognized and enacted. But data-rich literary history projects rarely consider how the disciplinary infrastructure analyzed relates to the historical context investigated. The lack of shared standards for data publication—and, more specifically, of a framework for combining these four features in investigating and representing the transmission and transformation of historical evidence to and in the present—problematises the field’s capacity to advance historical knowledge.

To remain with the projects introduced previously, neither So and Long, nor DeWitt, nor Cordell and Smith make anything like the same polemical claims to directly represent historical fact that Moretti and Jockers do; but none of these studies adequately investigate the relationship between the historical context they explore and the disciplinary infrastructure they use to conduct this exploration. So and Long
explicitly resist the view that their network model shows the evolution of modernist poetry: its nodes and edges are “an interpretation of the underlying publication data” (157). Yet in not assessing the quality of that data—or rather, in briefly noting limitations only with the bibliography they construct—So and Long do not consider the extent to which the interpretation they offer is determined by the scope and constitution of specific bibliographies.

While So and Long offer little information on data collection and construction, DeWitt extensively details this process for her study of genre formation, including the nineteenth-century periodical collections searched, the methods for extracting data from these results, and the size of the dataset gathered. Yet DeWitt similarly takes for granted that the resources she uses consistently and adequately represent the publishing context she investigates. DeWitt’s datasets are derived from searching the interfaces of large, proprietary collections, and she does not discuss the manner of their construction (for instance, the reliability of the OCR-rendered text or the effects of the search algorithms used) nor the extent to which their holdings represent Victorian literary criticism. The difficulty of conducting such an analysis of proprietary collections raises an important issue for data-rich literary history, and literary history in general: the adequacy, reliability, and transparency of emerging digital infrastructure. Proprietary mass-digitized collections such as Google Books, Early English Books Online, and The British Newspaper Archive (owned by Google, ProQuest, and findmypast, respectively) are increasingly used in humanities research. But their scope and scale—let alone the histories of transmission that produce them—can be very difficult to discern; indeed, the commercial imperatives of these enterprises arguably depend on them presenting these collections as comprehensive.

Yet even when researchers use nonproprietary mass-digitized collections and move beyond the search box to export and explore the underlying data, the scope and reliability of such holdings are not ascertained. As in DeWitt’s study, the Viral Texts project devotes considerable attention to describing data collection, including how its algorithm defines its object (reprinted passages), the characteristics of the resulting dataset, and the apparent and potential limitations of their method (Smith, Cordell, and Dillon; Smith, Cordell, and Mullen). While these authors note the inevitable difference between the
newspapers published in the past and the mass-digitized collections used to investigate reprinting, they do not characterize the nature and extent of these gaps. Despite the undoubted technical and critical sophistication of the Viral Texts project, and the valuable insights it has produced regarding American print culture, it does not explore the relationship of the mass-digitized collections it analyses—chiefly, the Library of Congress’s *Chronicling America*—to the historical contexts they are taken to represent. Even fundamental questions, such as approximately what proportion of American newspapers have been digitized by *Chronicling America*, are not addressed.

It may well be that the disciplinary infrastructure used by these projects adequately signifies the relationships of production and reception they seek to investigate, but this also might not be true, or it might be true to differing extents in the different cases. In not assessing whether and how well the data analyzed represents the historical context explored, these projects ultimately interpret the characteristics not of literary-historical systems but of particular components of our disciplinary infrastructure.

While the bibliographies and digital collections investigated by the above projects inevitably exclude multiple periodicals that existed in the past, the contemporary, born-digital collections that Finn investigates—for example, book recommendations and reviews on Amazon—embody the relationships of production and reception he seeks to explore. As well as describing the meaning of these relationships for literary culture in America, in explicitly defining his samples, Finn articulates a clear relationship between the collections he analyses and the samples he uses. For instance, with the Amazon “Customers Who Bought This Item Also Bought” category, Finn’s script identifies the first ten recommendations listed for a designated author, then repeats the process for two further iterations, to gather information on a further one hundred and then one thousand links. The resulting dataset is not the reviews of an author but “the three-level network surrounding a particular author’s work,” as represented by Amazon at a certain time (30).

While Finn’s account of data construction is exemplary, like So and Long and DeWitt, he does not publish the resulting dataset. The dataset that Finn uses to model the relationships of production and reception presented by Amazon—however representative at the time it was
constructed—became different to those relationships essentially from the moment he ceased harvesting data. Accordingly, no matter how detailed his information about data construction, the basis of his study cannot be reproduced: the version of Amazon that Finn explores no longer exists. Finn’s arguments refer to his own model, and with that model unpublished, no one else can engage with these arguments on the grounds on which they are made.

Data publication is integral to the Viral Texts project, which includes a searchable database of reprinted passages, including information on the periodicals encompassed by their study and locations and dates of publication for reprinted text. However, the form in which data publication occurs presents a related problem to Finn’s study. We are told that the project’s “database will never be a finished and polished archive, but instead an evolving and experimental space for exploring text reuse across an expanding set of corpora” (Smith, Cordell, and Mullen E14). On one level, this is an understandable response to working with (evolving and experimental) mass-digitized collections. But this approach means that the literary-historical arguments Cordell offers are not anchored in a stable and accessible dataset. The dataset he uses to explore reprinting in antebellum American newspapers, for instance, represents a stage in the development of analysis of a mass-digitized collection (“Reprinting”). With that expression subsumed into an “evolving and experimental” database, the basis of his argument no longer exists—at least, not in published form.

In chapter 1 I argued that Moretti’s and Jockers’s lack of data publication manifests and perpetuates their disregard for textual scholarship and the historical and interpretive meanings it offers. I do not think this is the reason for the problems with data publication—or lack thereof—in these other data-rich literary history projects. The extensive investment of time required to publish data and the lack of an established framework for doing so are the more likely culprits. But in not grounding their arguments in a stable, consistent, and freely available object of analysis, these projects manifest problematic consequences similar to those presented by distant reading and macroanalysis. Lack of access to the datasets used by these scholars makes it impossible for others to engage with their arguments in the terms in which they are made, or to reuse and repurpose the data. Because these datasets are so time-consuming to construct, lack of data publication is a
significant impediment to cumulative scholarly endeavor in data-rich literary history. Without access to underlying data, visualizations and other data summaries become the only available “text” for analysis. As well as concealing the data used, this situation occludes the fact that such summaries arise from—rather than simply invite—interpretation.

Most importantly, not publishing data obscures the fundamental nature of digital collections as “transactional objects whose very identity is constituted through exchange.” Drucker (“Distributed” 12) makes this point with respect to individual digital documents, but it is equally applicable to digital collections. Our inquiries to them—in conjunction with the protocols, conditions, and constraints that they enact and are enacted through—do not discover a preexisting version of the documentary past but configure and constitute one. Part of Drucker’s point is that this condition is not exclusive to the digital realm, and this also is true of collections: as I maintained above, analog collections have always been produced in transactional moments of production and reception. But the multiple, interactive, and shifting instances of representation and remediation constitutive of digital collections make this situation more acute and emphasize that the historical record does not exist independently of the structures and systems through which we access it. And as Drucker notes, there is “no way to preserve or recover the phantasm whose materiality is dependent on so many contingencies and co-dependencies of distributed hardware and related software, networks and clock speeds, protocols and display capacities” (26).

But there is—and there is only—the capacity to represent the effects of engagements with particular digital collections at particular times. Herein lies the fundamental importance of data publication for data-rich literary history: in expressing a materiality that no longer exists in any other form, it offers the only possible basis for conversation on shared premises. It is not enough to point to the mass-digitized collection or bibliographical database from which data were derived. The constitutive features of that entity have almost certainly altered. And even if the digital collection has not expanded (or contracted), data-rich literary history does not analyze the collection itself. It explores the effects of scholarly engagement with and interpretation of it.

Of the data-rich literary history projects discussed in this chapter, Underwood and Sellers’s work on changing standards of literary pres-
tiger most consistently enacts the curatorial elements present elsewhere in digital humanities, not least in terms of data publication. As the authors explain in an online working paper on the project, its most time-consuming element was not training their supervised model but constructing their dataset: identifying the different subgenres—poetry, prose, fiction, and drama—present in HathiTrust (“How,” 6). As well as publishing the datasets and code used in their article (“Code”), in collaboration with HathiTrust Underwood takes the major, additional step of releasing the outcome of analysis of that mass-digitized collection for others to use. This takes the form of “word counts for 101,948 volumes of fiction, 58,724 volumes of poetry, and 17,709 volumes of drama” published from 1700 to 1922, as well as yearly summaries of word frequencies for each genre. Underwood refers to this dataset as a “collection” to differentiate it from a “corpus” because “I don’t necessarily recommend that you use the whole thing, as is. The whole thing may or may not represent the sample you need for your research question” (“Dataset” np).

Signifying growing recognition of the importance of data publication, the Underwood/HathiTrust collection is an important undertaking for data-rich literary history in at least two ways. In presenting a dataset designed for literary history, it offers a shared foundation for research. Working with it, researchers can ask a range of questions based on a reliable, standardized dataset and engage with each other’s arguments in terms not only of results produced but of data investigated. In characterizing that collection as the holdings of “American university and public libraries, insofar as they were digitized in the year 2012 (when the project began),” Underwood also frames a major mass-digitized collection—HathiTrust—in terms of its history of transmission (“Dataset” np). However general this framing, Underwood thus explicitly associates the dataset he publishes with a sequence of production and reception that profoundly affects its capacity to support historical analysis. In their article, Underwood and Sellers acknowledge that this history of transmission shapes their findings, noting that their model “makes more accurate” predictions for American poetry collections because HathiTrust “mainly aggregates the collections of large American libraries” (“Longue” 338).

This approach to data description, curation, and publication recognizes that the relationship between a collection and a historical con-
text is never direct and transparent. But even Underwood and Sellers are equivocal in characterizing the broader relationship between literature in the past and the disciplinary infrastructure used to investigate it. Seeking to define the scope of their dataset, Underwood and Sellers 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.” Yet their own “work on fiction” with this collection belies the apparently solid basis of this estimate, finding that HathiTrust contains “many titles left out of” existing bibliographies (“How” n11). Underwood and Sellers thus indicate a significant lack of overlap between established bibliographical records and the holdings of a major digital library, although they do not highlight the significance of this finding nor explore its implications for their own study or for literary history broadly (whether conducted by computational or noncomputational means). Emphasizing that we cannot know the documentary past except through the knowledge infrastructure we create to interpret it, this disjunction that Underwood and Sellers discover highlights the potentially major gaps in all existing forms of description and interpretation: neither the analog nor the digital record offers an unmediated and comprehensive view of the documentary past; both are partial, and not necessarily in complementary ways.

III

The challenge 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 digitized documents and literary data it transmits)—suggests the strategy for meeting it. The scholarly edition has long offered both a theoretical basis and a practical technology for demonstrating and managing the documentary record’s partiality. Applied to the literary system rather than the literary work, the scholarly edition provides a framework for investigating the history of transmission constitutive of the literary system modeled, justifying the selections and decisions made in that analysis, and publishing the outcomes.

A conventional scholarly edition is not simply a version of a literary work. It is an “embodied argument about textual transmission,” as
Eggert puts it (Securing 177) or, in McGann’s words, a “hypothetical platform” for historical enquiry, one that both indicates and provides a pathway through the instability of the “textual condition” (“From” 203, 230). This dual capacity inheres in the interrelationship of the edition’s critical apparatus and curated text. The critical apparatus describes and justifies the editor’s engagement with the documentary record constitutive of the literary work, including the inevitable gaps and uncertainties that engagement exposes and creates. The curated text represents the outcome of that extended critical encounter, offering to those who accept its tenets—or who lack the expertise to engage with the documentary record in that manner—both an argument about the nature of the literary work and a foundation for analysis of that work.

For a scholarly edition of a literary system, the critical apparatus details the history of transmission by which the existence and interconnections of literary works in the past are known. Much more than simply describing the construction of a dataset—something already offered by many data-rich literary history projects—this critical apparatus elaborates the complex relationships between the historical context explored, the disciplinary infrastructure employed in investigating that context, the decisions and selections implicated in creating and remediating the collection or collections, and the transformations wrought by the editor’s extraction, construction, and analysis of that data.

A curated dataset replaces the curated text for a scholarly edition of a literary system. In the form of bibliographical and textual data, it manifests—demonstrates and, specifically, publishes—the outcome of the sequence of production and reception, including the current moment of interpretation, described in the critical apparatus. The model it provides is stable: it is published and accessible for all to use, whether for conventional or computational literary history. But that stability does not belie or extinguish its hypothetical character. Rather than showing a literary system, it presents an argument about the existence of literary works in the past based on the editor’s interpretation of the multiple transactions by which documentary evidence of the past is transmitted for the present. Its suitability and reliability for literary-historical research is established by a relationship between the historical phenomena and the data model that is explicitly interpretive and contingent rather than supposedly direct or natural.
As in a conventional scholarly edition, the nature of and issues foregrounded in any scholarly edition of a literary system will depend on what the editor or editorial team perceive as most relevant to understanding the existence of literary works in the past and the disciplinary infrastructure that evidences this. My scholarly edition of a literary system is profoundly influenced by the sociology of texts tradition (McKenzie). But this framework equally allows for modeling based on other theories of textual transmission. Whatever theoretical perspective informs it, where “the stylistic protocols of literary criticism” mean that issues deemed methodological are often relegated to footnotes or “methodological caveats”⁹—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 engagement with disciplinary infrastructure to effect the modeling of a literary system.

In its stability, this curated dataset resembles the Underwood/HathiTrust “collection,” and this feature is vital. Both publish the outcome of a critical encounter with digital disciplinary infrastructure and thus offer a consistent object for analysis that does not presume or pretend the stability of the documentary record. At the same time, a scholarly edition of a literary system differs from the word counts offered by the Underwood/HathiTrust collection, not only in providing a dedicated critical apparatus to define its construction 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 scholarly edition of a literary system should be accessible to all literary scholars through an interface for searching, browsing, and exporting the curated dataset.

Second, while the Underwood/HathiTrust collection is presented as raw data from which researchers can construct a sample, a scholarly edition of a literary system explicitly embodies an argument about historical relationships between literary works in a set time and place, at the same time as it offers a sample of literary works for analysis. In this sense, a scholarly edition is less potentially extensible than the Underwood/HathiTrust collection, although that collection also relates to a specific—albeit extensive—time period and collates a particular type and form of literary works. But a scholarly edition of a literary system
is not designed to be applicable to all times and places. Rather it seeks to provide an interpretive intermediary between increasingly complex and extensive digital disciplinary infrastructure and the requirements of literary-historical analysis.

As noted in the introduction, for the scholarly edition of extended fiction in nineteenth-century Australian newspapers that underpins this book, the critical apparatus is comprised of two parts: fields in the curated dataset, which detail decisions and arguments underpinning specific data constructions, and a historical introduction, presented in chapter 3 and delineating the principles of the modeled literary system in terms of the history of transmission by which it is constituted. The curated dataset is published in two sites and forms—as downloadable data alongside the Open Access version of this book on the University of Michigan Press website and in a database for searching, browsing, and selective or wholesale exporting of bibliographical and textual data.

This dual publication is an admittedly inelegant solution to the perpetual problems in digital humanities of sustainability and access. The first format is the most sustainable one I could devise, in that the data are held by the University of Michigan Library, in association with the press, and will be maintained by the significant resources of those institutions, I hope, well into the future. However, as with the Underwood/HathiTrust collection, in that form the curated dataset is usable only by scholars able to manipulate extensive textual and bibliographical data computationally. The database supports access for all literary scholars and permits me, and I hope others, to refer to fiction in the curated dataset in the knowledge that anyone can use the database to access the record’s full bibliographical and textual data. As I discuss in the book’s conclusion, the database also supports ongoing research into Australian newspaper fiction and the future utility of the data underpinning this book in providing facilities for users to interact with Trove to identify new fiction, enhance bibliographical information, and correct textual data. Yet conceivably, that format will be sustained only as long as I remain at the Australian National University. As digital humanities scholars work toward new forms of enhanced publication, I hope that a platform for both sustainable and accessible data publication will emerge; in the interim, this is the solution I have improvised.

For some researchers, a scholarly edition of a literary system will
function much as existing digital collections do: providing a site for searching or browsing the digitized documentary record. In this capacity, it also enables access to what could and I think should be a key contribution of data-rich literary history to the broader discipline: an expanded bibliographical record. While Underwood and Sellers note the presence in HathiTrust of new works not previously recorded by literary historians (who knows how many), the scholarly edition underpinning this project identifies new Australian titles and authors. For other researchers, including those currently using mass-digitized collections to locate individual authors and works in the historical context in which they operated, a scholarly edition of a literary system will offer a carefully, consistently, and—by the critical apparatus, explicitly—historicized digital collection for this task. And for researchers analyzing large-scale trends in the publication, circulation, reception, and contents of literary works, such an edition will provide a rigorously constructed and explained “shared” dataset, that could be analyzed and “combined in more ways than one” (Moretti, *Graphs* 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 nuance of such analyses foundationally depends on the historical knowledge embedded in those constructions. With a scholarly edition of a literary system already, in and of itself, an argument about a “collective system . . . as a whole” (Moretti, *Graphs* 4), analyses of it can attend to the multiple features and dimensions it models. Alan Liu’s notion of “contingency” encapsulates the resulting analytic mode. Describing the relationship of historicism to the database, Liu notes that in neither form does one ask what is this whole, what does it mean? Rather, the question becomes how does this complex system I am investigating appear from this perspective? Although the whole is composed of parts that are explicitly defined and constrained—“chained to context,” as Liu puts it—the mode of inquiry it supports is transactional, enabling unpredictable connections and insights (*Local* 262).

As with a scholarly edition of a literary work, a scholarly edition of a literary system is thus intended not to conclude but to support various forms of investigation, including those that move between the single literary work and the system in which it existed and operated, as I demonstrate in chapter 6. The issues explored in the book’s second half
reflect my interests, and the topics I believe are important in understanding nineteenth-century literary culture, in Australia and globally. But as I show in the conclusion, these chapters in no way exhaust—or come close to exhausting—the potential questions and arguments that this scholarly edition of a literary system might support, whether enacted through traditional or computational means.

The approach to data-rich literary history I am advocating does not take the path increasingly recommended for the field: of integrating scientific and social scientific measures of statistical uncertainty into historical analysis (Goldstone, “Distant”). Given that constructing 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 to 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 a critical apparatus and curated dataset, the framework of the scholarly edition offers a theoretical and practical basis to model the relationships of production and reception that constitute historical literary systems, while assessing and managing the inevitable contingency of those relationships and of the documentary infrastructure through which we perceive them. As well as supplying a dedicated object of analysis for data-rich literary history, a scholarly edition of a literary system seeks to extend the insights gained from that field’s engagement with emerging digital disciplinary infrastructure to the broader discipline. Conducted on that basis, data-rich literary history could transform from an unexpected, often unwelcome intruder into a vital interlocutor between literary history and the digital context in which it increasingly operates.