TIM SOMMER

Canon, Corpus, Archive:
Selection and Valuation from the 18th Century
to the Digital Humanities

1. Introduction

In his essay "Conjectures on World Literature" (2000), Franco Moretti reminds his readers that "there are thirty thousand nineteenth-century British novels out there, forty, fifty, sixty thousand," before pointing to the essential futility of simply "[r]eading 'more'" – a small-scale, short-term remedy but ultimately "not the solution" (2000a, 55) to the larger problem of not-reading. Proposing what he influentially termed "distant reading" as a viable alternative, Moretti gestures towards a methodological optimism characteristic of much work in the digital humanities since the original appearance of the essay two decades ago. A few paragraphs later, he more specifically addresses the relationship between canonicity and the type of approach his "Conjectures" seek to overcome: "the trouble with close reading," he points out, "is that it necessarily depends on an extremely small canon. [...] [Y]ou invest so much in individual texts only if you think that very few of them really matter" (ibid., 57; original emphasis). A similar critique of the limited critical reception of a small body of texts surfaces in another one of Moretti's turn-of-the-century essays, "The Slaughterhouse of Literature." Victorian fiction here again forms the basis for a diagnosis of the shortcomings – both quantitative and qualitative – of traditional ways of reading: "The majority of books disappear forever – and 'majority' actually misses the point: if we set today's canon of nineteenth-century British novels at two hundred titles (which is a very high figure), they would still be only about 0.5 percent of all published novels" (Moretti 2000b, 207; original emphasis). Moretti hastens to clarify that what may sound like a return to the logic of the late-20th-century canon wars is in fact an argument that transcends earlier debates about inclusion and exclusion. The "aim" of distant reading, he suggests, "is not so much a change in the canon [...] as a change in how we look at all of literary history: canonical and noncanonical: together" (ibid., 207-208; original emphasis). The ultimate objective, in other words, is not to render the canon more inclusive, but to do away with the very idea of canonicity itself through aspiring to account for what Moretti, drawing on Margaret Cohen, calls "the great unread" (ibid., 208).

While Moretti has been an early champion of such an ambitious programme, he has hardly been the only one to employ its underlying reasoning in order to highlight the potential of digital scholarship. Matthew Jockers, who speaks of 'macroanalysis' rather than 'distant reading,' offers a similar picture of the transition from evaluative selection to reading the entirety of the literary record:

The existence of huge data sets means that many areas of research are no longer dependent upon [...] observations derived from data sampling. Instead of conducting controlled experiments on samples and then extrapolating from the specific to the general...
or from the close to the distant, these massive data sets are allowing for investigations at a scale that reaches or approaches a point of being comprehensive. (2013, 7)

Jockers's language is more technical than Moretti's, but the overall thrust of his argument is quite similar. Literary studies have failed to seize on the opportunities provided through such progress, Jockers suggests, given that "the scaling of our scholarly questions" dramatically lags behind "the massive scaling of digital content that is now held in twenty-first-century digital libraries" (ibid., 16-17).

What emerges from statements such as these is a progressivist disciplinary history of literary studies that juxtaposes a narrowly canon-reliant past with a not-too-distant future in which the canon will have become replaced by large-scale corpora that quasi-objectively map literary history as a whole. This contrast conceptualises the dialectical development of literary studies as what James English and Ted Underwood have described as "a drama of competing scales" (2016, 278). The negative image that many contemporary visions of big-data analytics use as a contrastive foil is the impressionistic kind of selection and valuation they depict as having dominated literary criticism and scholarship in the past – exemplified, for instance, by the idiosyncratic preferences of Romantic poet-critics like Wordsworth and Coleridge, by Matthew Arnold's celebration of "the best that has been thought and known" (1993, 79), or by the latter-day Romantic Harold Bloom's defence of the "Western canon" (1994). The historical and methodological development that theoretical digital humanities writing often evokes is one from canon to corpus to archive – terms that the following two diagrams visualise as relationships between supersets and subsets.

Figure 1: The relationship between canon, corpus, and archive as frequently conceptualised in digital humanities scholarship.

The first diagram (figure 1) illustrates the link between the terms as Moretti, Jockers, and others present it: a tiny, clearly circumscribed, and essentially stable canonical area at the centre; the digital corpus as an exponentially larger range of texts enveloping and including the canon; and the archive (or "all of literary history") as the most comprehensive scale of literary analysis – a historical totality that the digital...
The corpus is imagined as ultimately capable of matching. The following diagram (figure 2) provides a more complex visualisation of the terms that maps a range of different archives and corpora:

![Diagram of archives and corpora](image)

Figure 2: The relationship between canon, different types of archive, and different types of corpus.

"All of literary history" here denotes everything ever written. Only parts of this have been published and of these texts, in turn, only a smaller subset has been preserved in the analogue archive (libraries, private collections, and other repositories). The digital archive – which in the first diagram is almost identical with literary production as such – here emerges as a more limited subset of the whole, several distances removed from that potential entirety. As will become clear from my analysis of individual digital projects below, the actual corpora that distant-reading or macroanalytical research uses as the basis for its computational processing of literary data are selections sampled not from the whole, but from a smaller subset. In some cases, corpora are designed to replicate the digitised non-canonical archive ("Corpus 1" in figure 2), in others they overlap ("Corpus 2") or are wholly contained within the canon ("Corpus 3"). Due to factors including copyright restrictions on 20th- and 21st-century writing, the digital archive technically speaking does not even contain the entirety of the canon (which in figure 2 hence extends beyond the boundaries of the digitised).

My larger point here is to highlight that digital corpora are not objective givens, but the result of deliberate choices – of processes of selection (which texts should be chosen, which should be excluded?) and valuation (which dataset furnishes the best random sampling of the digital archive as a whole, which provides the best basis for a...
specific research question?). The digital archive, moreover, is hardly a neutral representation of the totality of literary production. As Katherine Bode has recently pointed out, the texts featured in it "are only some of those that existed and some of those that were collected by an institution" (2018, 44). Canonicity, in other words, is a built-in feature of the digital archive as well. Digitisation does not so much solve the problem of selection as reintroduce it on a different scale. One of the questions that this raises is from what distance distant reading actually operates. There is a qualitative (that is, a material and technological) difference between human reading and computational analysis, but not necessarily a quantitative difference in terms of the number of texts under consideration – certainly not, at least, a difference that corresponds to a ratio of 99.5% to 0.5%.

In order to make sense of this fact – and of the maximalist language by which digital research has often been accompanied –, it is useful to reconstruct what I have referred to above as the negative image of canonical reading and criticism. What follows throws a series of spotlights on different manifestations of literary selectivity. I will first look at the criteria pre-Romantic and Romantic writers – authors, critics, editors, and anthologists – employed in choosing and evaluating the significance of texts before I move on to some specific examples of corpus building and textual selection in the digital humanities.1 My main objective will be to substantiate two related claims: first, that the historical period commonly associated with narrow hyper-canonisation was marked by a simultaneous turn towards the comprehensiveness of the archival; and, second, that, just as there is a maximalist countertendency to canonisation as early as the late 18th century, the minimalism of curation and selective sampling in turn continues to play a constitutive role in contemporary digital scholarship.

2. Canon, Corpus, Archive, c. 1750-1950

To shed light on what has often been described as the age of the invention of canonicity, it is useful to look at the criteria employed in the late 18th and early 19th centuries to determine which texts qualify for preservation and interpretation. Some of the major topics that emerge in this context include the following: elegance or wit (mainly an 18th-century criterion); nationality or national representativity (a subject that becomes increasingly important over the course of the 19th century); disinterestedness and anti-commercialism (an authorial more than a textual property, but a criterion that has been crucial from Romantic writing onwards); and formal sophistication or experimentalism (something that peaks with the Modernist poet-critics but has a strong presence both before and after Modernism). These criteria are often overlapping (formal experimentalism and a low market profile are frequently linked, for example), but there

---

1 The label 'digital humanities' has, of course, become an umbrella term for a wide range of topics and methodologies, from new media studies to big-data projects in fields such as history or sociology. I am here concerned more specifically with the computational analysis of literary texts.
is also, at times, a tension between them (for instance between the criterion of national representativity and that of stylistic sophistication).

Anthologies are a genre that is especially helpful to reconstruct how these criteria have historically been applied. They provide tangible evidence of what has been deemed worthy of inclusion, but they also often feature explicit editorial commentary on the motives behind the selection of some texts over others. In William Hazlitt’s preface to his *Select Poets of Great Britain* (1825), for example, the idea of national value takes centre stage, illustrating the close relationship between literary canonisation and cultural nationalism. Describing his ‘ambition’ as the ‘compiler’ of the 600-page collection of British verse, Hazlitt writes that he aimed
to offer to the public a *body of English poetry* [...] such as might at once satisfy individual curiosity and justify our national pride. We have reason to boast of the genius of our country for poetry and of the trophies earned in that way; and it is well to have a collection of such examples of excellence inwoven [sic] together as may serve to nourish our own taste and love for the sublime or beautiful, and also to silence the objections of foreigners, who are too ready to treat us as behindhand with themselves in all that relates to the arts of refinement and elegance. [...] Poetry is one of those departments in which we possess a decided and as it were natural pre-eminence: and therefore no pains should be spared in selecting and setting off to advantage the different proofs and vouchers of it. (1825, v; original emphasis)

Hazlitt leaves no doubt about the fact that his editorial effort has been one of "selecting" and arranging: he offers "a body of English poetry," not English poetic production as a whole. Although his sample of this larger archive is taken from literary history from the Middle Ages to the then near present (with texts ranging from Chaucer to Robert Burns), Hazlitt is remarkably frank about the necessity of cherry-picking from this totality in order to make a convincing – and internationally recognised – case for English literary merit.

In the first half of the 19th century, anthologising becomes an even more explicitly nationalist project on the other side of the Atlantic. In the prefatory remarks to his 1842 collection *The Poets and Poetry of America*, Rufus Wilmot Griswold, like Hazlitt before him, describes his anthology as an effort to gather evidence of national literary excellence:

This work is designed to exhibit the progress and condition of Poetry in the United States. It contains selections from a large number of authors, all of whom have lived in the brief period which has elapsed since the establishment of the national government. Considering the youth of the country, and the many circumstances which have had a tendency to retard the advancement of letters, it speaks well for the past and present, and cheeringly for the future. (1842, v)

Yet in contrast to Hazlitt, who emphasises the extent of his selection activity, Griswold is forced to admit that, surveying the field, he was not confronted with an embarrassment of riches. Compelled to settle for an aesthetic compromise in order to launch his project of advertising American literature, he privileges the criterion of national representativity over that of formal achievement:
In selecting the specimens in this work, I have regarded humorous and other rhythmical compositions, not without merit in their way, as poetry, though they possess but few of its true elements. So many mistake the form for the divine essence itself, that I might have experienced difficulty in filling so large a volume, had I been governed by a more strict definition. (ibid., vi)

Where Hazlitt highlights selectivity, Griswold is motivated by the idea of encyclopaedic coverage. If the former explicitly supplies a canon of 'select poets,' the latter aspires to represent the national poetic archive as a whole.

But Griswold in this respect appears to be the exception rather than the rule. There are countless examples of a 19th-century celebration not of the holism of the archive, but of the urgency of selection. Wordsworth, for instance, at one point suggests that the majority of texts deserves being consigned to the dust-heap of literary history – that one of the major "errors, into which we easily slip when thinking of past times [...] lies in forgetting, in the excellence of what remains, the large overbalance of worthlessness that has been swept away" (1974, vol. 2, 9). In the 1800 "Preface" to Lyrical Ballads, he similarly reflects on the difference between shallowness and depth, juxtaposing the eminence of the purified old canon with the aesthetic deficiencies of contemporary mass literary production – applying a combination of the complexity and the anti-commercial criterion. With Wordsworth, as John Guillory has argued, "the prestige of literary works as cultural capital is" beginning to be "assessed according to the limit of their dissemination, their relative exclusivity" (1993, 133; original emphasis) – a process that Pierre Bourdieu, writing about French literary culture, has described as a "division between the field of restricted production [...] and the field of large-scale production" (1993, 39).

The same differentiation between rarefied 'writers' writing' and literary production at large appears in a number of Victorian authors. The long afterlife of this rationale of selection can be gauged from a text such as F. R. Leavis's The Great Tradition (1948), which is perhaps best remembered for defining a narrow hyper-canon of "great English novelists" (consisting of Jane Austen, George Eliot, Henry James, and Joseph Conrad) (1955, 1). Leavis excludes Charles Dickens, for example, on the grounds that he was "a great entertainer" who "had for the most part no profounder responsibility as a creative artist than this description suggests" (ibid., 19). But Leavis also, on a more abstract level, reflects on the difference between the archive – Moretti's "all of literary

---

2 "The invaluable works of our elder writers," Wordsworth fears, "are driven into neglect by frantic novels, sickly and stupid German Tragedies, and deluges of idle and extravagant stories in verse" (1974, vol. 1, 128). This type of argument obviously surfaces in many writers of the period. It also, for example, informs Griswold's guilt-ridden admission of having had to lower his aesthetic standards for political and commercial reasons.

3 Thomas Carlyle, for instance, sternly criticised Walter Scott for the commercial orientation and profuse extent of his writing: "productive without measure as to quantity, in quality he for the most part transcended but a little way the region of commonplace" (Carlyle 1872, 33). On the underlying late-Romantic poetics that manifests itself in this critique, see Sommer (2017, 367-370).
history” — and the canon, making a case for the importance of the latter when he argues that "far from all of the names in the literary histories really belong to the realm of significant creative achievement" (ibid., 2). Leavis's is a 20th-century version of Wordsworth's "overbalance of worthlessness" argument, opposed — like its Romantic precursor — to attempts at a holistic recovery of the archive.

Although this kind of selection bias proved extremely influential, it is nevertheless worth emphasising that there was also — both before and contemporaneous with Wordsworth — an argument against the limitations of canonicity and instead for something like the totality of the archive. This can be seen, for instance, in the lively late-18th-century interest in collected editions of the English poetic tradition. Among the large number of these projects were London bookseller John Bell's *The Poets of Great Britain Complete from Chaucer to Churchill*, which appeared in 109 volumes between 1776 and 1782; the 68-volume *The Works of the English Poets; With Prefaces Biographical and Critical*, published between 1779 and 1781 and perhaps the most famous of these editions because it contained 'Lives' of the poets written by Samuel Johnson; and the 124-volume *Poets of Great Britain from the Time of Chaucer to Sir William Jones*, edited by Samuel Bagster and published in 1807. 4 Rather than providing the 'best' — or the critically most acclaimed — selection (as in the case of Hazlitt's anthology), late-18th-century publishers and editors were competing with each other to offer ever larger and more comprehensive editions. A collection that illustrates this editorial maximalism particularly well is Robert Anderson's *The Works of the British Poets, with Prefaces Biographical and Critical*, which also ran under the more suggestive title *A Complete Edition of the Poets of Great Britain*. In his preface to the venture, Anderson criticises his precursors for having been too narrow in their selections. He himself confidently promises to offer "a Collection of English poetry, which contains the works of one hundred and fourteen authors, of whom forty-nine are not to be found in [...] Dr. Johnson's edition; and forty-five are now, for the first time, received into an edition of English poetry" (1795, 5; original emphases). Like his colleagues — and like some of his 21st-century digital humanist successors —, Anderson is obsessed with the sheer metrics of literary history.

Like the analytical maximalism championed by Moretti, Jockers, and others, such 18th-century canon-broadening endeavours were the result of technological innovations as well as of a disciplinary paradigm shift — in this case revolutions in printing and in paper production that created a cheap supply for a fast-growing reading public (see St Clair 2004) and, on the other hand, a booming antiquarian recovery of the English literary past. It was partly this mass-availability of print products present and past that drove Romantic literary criticism into creating canonical scarcity out of a new kind of archival abundance. On closer inspection, however, even a writer like Wordsworth turns out to be less of a minimalist than he might at first appear. When he wrote to Anderson in September 1814, some twenty years after the appearance of the latter's edition, it was not to criticise him for having been too ample in his selection, but to accuse him of narrowness. While he admitted that readers were "much indebted to

---

4 For a detailed book-historical account of the cultural background of these collected editions and the economic rationale behind printing them, see Bonnell (2008).
Anderson "for the, at that time, unexampled comprehensiveness" of his edition, Wordsworth regretted to say that the British "Public" was "still unprovided with an entire Body of English Poetry" (1970, 151-152). "I have long wished this to be done," he continued, "and have talked with several of my Friends [...] upon the subject who [...] participate in my desire to see your Edition adequately enlarged" (ibid., 152). Wordsworth here deliberately imagines "an entire body" – literally, a complete corpus – of poetic writing rather than a more narrow canon; he ultimately wants to "enlarge" instead of winnowing down. What is on display in both the paratextual framing and the critical reception of collections of the kind edited by Anderson is a striking preoccupation with numbers, one that manifests itself in the attempt to recover an ever larger part of the literary past and make it available to an ever larger readership.5

3. Selection and Valuation in the Digital Humanities

Distant reading has more recently revived this accretionist tradition and, in so doing, has reignited the critical debate about literary canonicity. Although digital humanities practitioners often stress that they are not interested in the type of enfranchisement activism championed by late-20th-century reformers of the canon, their progressivist vocabulary nevertheless implies that a corpus that comprises "all of literary history" will ultimately create conditions of equality and undermine the hegemonic authority of previous generations of critics (and that of their necessarily limited selections). The notion that digital methods will solve the problem of canonicity has been accompanied by scepticism – both from within and from outside of the digital humanities – about the pitfalls of technology's "lure of objectivity" (Rieder and Röhle 2012, 71). Taking issue with Jockers's 'macroanalytical' approach, Barbara Herrnstein Smith, for example, has cautioned against the idea "that computational methods and big data allow researchers to avoid or mitigate the subjectivity and bias built into human observation" (2016, 67). In a 2012 essay on "Race and the New Digital Humanities Canon," Amy Earhart highlighted the disappointingly negligible effect that digital approaches have had on the size and shape of the literary canon. Familiar arguments about inclusion and exclusion resurface when Earhart observes that "the canon that we, as digital humanists, are constructing" is one "that skews toward traditional texts and excludes crucial work by women, people of color, and the GLBTQ community" (2012, 316).6 Taken together,

5 If these editions illustrate an encyclopaedic ambition to assemble a complete record of the literary past, in another sense they are, of course, also instances of canonisation – both in terms of genre (they compile poetry as a high-prestige form, but not, for instance, fictional prose) and in terms of literary nationality (to the extent that they establish a vernacular poetic tradition worthy of preservation). Following this line of argument, Michael Gamer has recently suggested that Romantic writers' engagements with these editions "eerily anticipate the late twentieth-century debates over the canon" (2017, 18).

6 There has also been a more radical type of critique directed at the purportedly 'neoliberal' spirit at work in the digital humanities. This sees digital scholarship not as a quantitative leap beyond the faultlines of the canon debate (as Moretti would have it) but as a reactionary step backwards that threatens to nullify the achievements of canon revisionism. In an essay on

Anglistik, Volume 31 (2020), Issue 2
© 2020 Universitätsverlag WINTER GmbH Heidelberg
such responses suggest that digitisation has led not to a decline, but to a resurgence of canon discourse, one in which digital humanists themselves – despite, or perhaps precisely because of their professions of quantitative disinterestedness – have come under fire as champions of exclusionist selectivity.

The latter is not, however, the argument I wish to advance here. What I want to draw attention to lies at a more basic analytical level and concerns the fact that the digital humanities have not rendered selection and valuation as such obsolete. In this section, I look at two of the "pamphlets" published by Stanford University's Literary Lab to illustrate the extent to which the necessity to select texts remains a methodological challenge in digital scholarship. The Literary Lab was founded by Moretti and Jockers in 2010, and the working papers of the group have frequently dealt with the question of canonicity. Pamphlet 8, Mark Algee-Hewitt and Mark McGurl's "Between Canon and Corpus: Six Perspectives on 20th-Century Novels" (2015), revolves around an attempt to build a digital corpus of 20th-century English-language fiction. In the style of a research diary, the text charts the various "dilemmas of selection" the authors encounter in the course of this endeavour (2015, 1). Early on, they realise that they will not be able to represent the whole of the archive, but only a small, more easily manageable slice of the whole: a total of 350 texts. The dilemma arises from having to decide which 350 texts to choose for the corpus. Algee-Hewitt and McGurl are troubled by the idea of having to work with a random sample drawn from the entirety of the archive. "A corpus so constructed," they fear, "might suffer from a sense of mere arbitrariness, leaving out too many things – including most of the individual authors, certainly, and perhaps also whole genres and long phases of development – that scholars have come to care about" (ibid., 3; original emphasis).

what they call the "Political History of Digital Humanities," Daniel Allington, Sarah Brouillette, and David Golumbia, for instance, claim that "despite the aggressive promotion of Digital Humanities as a radical insurgency, its institutional success has for the most part involved the displacement of politically progressive humanities scholarship and activism" (2016, n.p.). To them, the super-human scale involved in digital archive-building figures not as a solution to the problem of canonical exclusivity, but as a reactionary totalitarianism that renders the subversive potential of the inclusion of marginalised literary voices statistically insignificant. For an eloquent rebuttal of wholesale critiques of the digital humanities (which have alternately portrayed the approach as reactionary, ahistorical, anti-theoretical, or economically complicit), see Matthew Kirschenbaum's 2014 essay "What is 'Digital Humanities,' and Why Are They Saying Such Terrible Things about It?" Kirschenbaum suggests that a fruitful critical engagement with work in the field depends on paying close attention to specific projects rather than reading the digital humanities (and the rhetoric that surrounds it) as an abstract whole – an assumption that to some extent paradoxically reinvests in the virtues of close reading. My own argument in the following proceeds from an analysis of two case studies to draw larger methodological inferences.


8 Several of the pamphlets – 17 of them have been published to date – were collected as Canon/Archive: Studies in Quantitative Formalism from the Stanford Literary Lab (Moretti 2017).
Driven by a residual yearning for the canonical, the two decide to create a more deliberately devised corpus that replicates the traditional canon rather than the featureless totality of the archive – explaining that a "randomized selection of books might function as a viable corpus, [...] but we found ourselves thirsting, after all – and ominously – for a body of data that would have at least some of the attributes of a *canon*" (ibid., 3-4; original emphases). This is obviously tongue-in-cheek, but the remark points towards a larger tendency in digital humanities research to legitimate its methodology through generating insights into texts that a wider scholarly and critical community recognises as culturally relevant and historically significant.  

Algee-Hewitt and McGurl seem hardly troubled by this attachment to the canonical, however. Once they have firmly left behind the aim of coming up with a random sample, the question becomes which authority to follow in deciding which texts count as the most canonical representation of 20th-century Anglophone fiction. Algee-Hewitt and McGurl collect five different lists of one hundred novels each – one of them representing critical prestige, another one providing the titles of annual bestsellers, and so on. As they soon come to realise, however, the canon that results from the superimposition of these lists is hardly inclusive in terms of the gender and ethnic backgrounds of the authors featured in it – an observation that leads Algee-Hewitt and McGurl to contact "specialists in ethnic or feminist literatures" (ibid., 15) for their additional recommendations. The corpus the pamphlet ends up assembling thus closely mirrors the interventions of canon war revisionism, but while this allows Algee-Hewitt and McGurl to deflect the kind of criticism voiced by Earhart and others, it leaves the canon/archive dilemma essentially unresolved.

Authored by Algee-Hewitt, Moretti, and four Stanford graduate student collaborators, pamphlet 11, "Canon/Archive: Large-Scale Dynamics in the Literary Field" (2016), opens with a more assertive appropriation of maximalist rhetoric. "Of the novelties introduced by digitisation in the study of literature," the authors suggest, "the size of the archive is probably the most dramatic: we used to work on a couple of hundred nineteenth-century novels, and now we can analyze thousands of them, tens of thousands, tomorrow hundreds of thousands" (2016, 1). This prospective optimism is followed by an extended definition of three key terms that appear throughout the pamphlet – the "published," the "archive," and the "corpus:"

---

9 Matthew Wilkens, in an essay on "Canons, Close Reading, and the Evolution of Method," associates this kind of canon-affirmation especially with what he calls "first-wave digital projects" – research that he criticises for having worked on the same basic premise as traditional literary scholarship, "the assumption that we already know which authors and texts are the ones to which we should devote our scarce resources" (2012, n.p.).

10 Indeed, as Timothy Aubry observes, their methodological choices are indicative not only of "their loyalty to precisely the nonscientific methods of inquiry and evaluation that the digital humanities has at times seemed dedicated to replacing," but also of their implicit acknowledgement "that the digital humanities represents a systematic evacuation of 'valuation'" (2018, 214).
[The idea of the published] is simple: it's the totality of the books that have been published (the plays that have been acted, the poems that have been recited, and so on). [...] The archive is [...] that portion of published literature that has been preserved – in libraries and elsewhere – and that is now being increasingly digitized. The corpus, finally, is that portion of the archive that is selected, for one reason or another, in order to pursue a specific research project. The corpus is thus smaller than the archive, which is smaller than the published: like three Russian dolls, fitting neatly into one another. But with digital technology, the relationship between the three layers has changed: 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.

When we use the term "archive," what we have in mind is precisely this potential convergence of the three layers into one; into that "total history of literature" [...] that used to be a mirage, and may soon be reality. (ibid., 2)

Comparing these definitions and the visualisation provided in figure 2 reveals some important dimensions of the relationship between canon, corpus, and archive that the pamphlet silently elides. The crucial distinction between the physical and the digital archive is largely absent from the above definition and the term 'corpus' is understood as a large-scale entity rather than as a more modestly sampled fraction of the archive. Moreover, the promise of an imminent convergence of these sets and subsets misrepresents actual proportions by suggesting that the archive in its entirety is on the verge of becoming available and readily legible.

The initial aim of the project traced in the pamphlet was to construct a corpus that functions as a reliable approximation of the archive and then to compare this body of texts computationally with another corpus that represents the canon. Thus, from the very outset, the idea of working with the entirety of the digitised literary record makes way for a more pragmatic approach that relies on a process of selective sampling. One way of arriving at the archive-corpus would have been to assemble a random collection of texts already available in digital databases. But these, as pointed out above, incline towards the canonical, their relative size notwithstanding. Openly acknowledging this problem and committed to working against this built-in bias, the authors of the pamphlet set out to create a genuinely randomised sample of the archive. The methodological path they choose to pursue is equally problematic, however, to the extent that the list from which their sample is derived is drawn from corpus types that are the result of curation as well (short title catalogues and bibliographies).

The "random sample" (ibid., 2) with which Algee-Hewitt, Moretti, and their co-authors end up is a list of 674 novel titles. They quickly discover that the more obscure items from this sample are hard to get hold of, either in digitised form or as physical copies. Only 30% of the novels in this archive-corpus are readily available through institutional subscriptions, another 40-50% turns out to be part of other digital databases. But it is the lack of the remaining titles (some of them apparently not held by any library worldwide, most of them too expensive to digitise) that is truly problematic, as the authors themselves concede. The fact that "the missing 20-30% would be, almost by definition, furthest from all conceivable forms of canonisation" threatens to render their "findings questionable" (ibid., 3). Their failed experiment leaves them with the sobering realisation that "the idea that digitisation has made
everything available [...] is a myth” (ibid., 3). In the end, they decide to compromise and draw a simpler sample from the available online collections. They then proceed to build a corpus representative of the canon and compare the two computationally for markers such as word redundancy and the co-occurrence of word-pairs, drawing larger inferences about the differences between canon and archive as though they had in fact managed to come up with something approaching literary history as a whole.

It is only in the conclusion to the pamphlet that the group acknowledges that this procedure represents a drifting away from the initial research question – a trajectory similar to the one that emerges from Algee-Hewitt and McGurl's project. Indeed, this discrepancy between aims and results is a characteristic feature of the Literary Lab pamphlets as a genre. As Ted Underwood has pointed out, the texts "are unified by a deliberately wandering structure, which keeps its distance both from scientists' predictable sequences (methods → results → conclusions), and from the thesis-driven template that prevails in the humanities (counter-intuitive claim → evidence → I was right after all)” (2017, n.p.). If what Underwood describes as the "progressive disorientation" (ibid., n.p.) that the pamphlets dramatise presents readers with a suspense-driven explorative approach, it also engenders a striking contrast between an initial commitment to a large-scale claim and the eventual frustration of the expectations fuelled by that claim. This structure also informs a text like Moretti's "Conjectures," which describes its own methodology as akin to "the form of an experiment. You define a unit of analysis [...] and then follow its metamorphoses in a variety of environments – until, ideally, all of literary history becomes a long chain of related experiments" (2000a, 61-62; original emphases). But Moretti candidly qualifies the potential of this procedure in a footnote appended to the passage: "How to set up a reliable sample," he confesses, "is of course quite a complex issue" (ibid., 62). In the small print, his contribution emerges as merely a "preliminary sketch" whose "sample (and its justification) leave much to be desired" (ibid., 62).

This disconnect between a notionally available totality of literary production and the problems that arise from its methodological implementation demonstrates the persistence of selection and valuation as forms of critical practice. The point of both pamphlet 11 and pamphlet 8 eventually becomes not to say something about the archive...
itself, but to draw inferences about the contrast between the canon and a smaller-scale representation of the archive. Instead of doing away with selection and valuation, research projects like the ones discussed here openly embrace both activities, given the amount of attention and painstaking labour they invest in devising, building, and curating digital corpora for subsequent analysis.

4. Conclusion

This pragmatic appeal of the subset – as opposed to the vertiginous complexity of the whole – looks familiar if seen against the background of the type of canon-making selection that is often described as a soon to be superseded way of conceptualising literary history. Given that past notions of minimalist hyper-canonicity were dialectically related to their maximalist counterpart, what one registers across the larger historical trajectory of the past 250 years is thus precisely not a straightforward development from narrow Romantic – or pre-Romantic – canons via their late-20th-century inclusionist broadening towards a convergence of digital corpora with literature in total. Highlighting close reading's limited ability to comment on literary history at large, digital approaches have thrown into relief what Andrew Piper describes as "a metonymical crisis [...] at the core of literary criticism," a methodological crux that revolves around "an incommensurable relationship between part and whole" (2018, 7).

In this larger debate about scale and method, quantitative work has tended to advance a post-mereological argument that suggests that digitisation is – or will soon be – capable of replacing partial knowledge with a knowable whole. Where data selection and corpus building continue to be elements in the process of knowledge production, however, the problem of generalisability remains a challenge.

Two recent accounts of distant reading have put such questions of scale centre stage. In the preface to his *Distant Horizons: Digital Evidence and Literary Change* (2019), Ted Underwood offers a digital humanist's retrospective account of the transformations of the approach over the past two decades. He concedes that – contrary to much manifesto language – the archive itself has often not been the main concern. While distant reading "was initially understood as an extension of the canon-expanding recovery projects of the 1990s," the actual "point" of the method, he suggests, "is not to recover a complete archive of all published works but to understand the contrast between samples drawn from different periods or social contexts" (2019, xx). Selection to Underwood remains a constitutive part of literary scholarship in the digital age, but its nature shifts from aesthetic valuation to representative sampling. In the case of the "Canon/Archive" pamphlet discussed above, this is no doubt an apt description of the rationale behind the selection process at work in the project – a method of textual assembly that differs from the criteria applied by a writer like Hazlitt. But as Algee-Hewitt and McGurl's "Between Canon and Corpus" demonstrates, aesthetic valuation continues to inform sampling practices in computational scholarship, just as the Romantic period witnessed the coexistence of the conflicting impulses of canonisation and archivisation.

If Underwood presents a pragmatic account of the nature and aims of distant reading, a different kind of conclusion emerges from a recent essay by Nan Da, who...
has raised doubts about the heuristic value of the analysis of large datasets. Pointing to the effort involved in building and analysing such digital collections, she suggests that computational work does not, after all, extend the limits – that is to say, the mere quantity – of what literary scholars and critics are already able to read in an analogue fashion. She here differs from Underwood, who argues – quite plausibly, I think – that computational analysis is able to yield something else than 'mere' reading, no matter how small a given sample may be. But whether that result is of larger interest is of course a different question. Da clearly thinks that it is not, concluding her essay by suggesting that although there may be "lots of questions we haven't answered [...] [t]hat does not mean that [...] any answer to any previously unasked question, or any question, is automatically worthy of attention" (2019, 639). If this is reminiscent of Wordsworth's wholesale dismissal of "the large overbalance of worthlessness that has been swept away" (1974, vol. 2, 9), it also shows that the discourse of canon, corpus, and archive ultimately raises not just the question of scale, but also that of relevance (whether of texts or of critical arguments). Da shares with Wordsworth a sense of the expedience of limitation, as much as Moretti and Jockers share with 18th-century anthologists a desire for comprehensive coverage. Seen from such a longue durée perspective, the line that separates past from present responses to the challenge of sampling is less clear-cut than we often tend to assume.

Works Cited


In a response to Da's essay, Piper appropriately enough uses a scale argument to question the computational validity and larger relevance of her findings. Where Da is sceptical about the value of the insight granted by distant reading, Piper contends that her own methodology suffers from an overly narrow selection of studies to advance her case and ultimately points to the larger "problem of selective reading" (2019, n.p.) that Piper sees as underlying work in the humanities in general (whether digital or analogue).


