

LAWRENCE EVALYN
Dissertation Committee Meeting, 20 May 2020



Emily Friedman (Auburn U) = c18 and DH Person → Director of 18thConnect.org = an aggregation site that peer reviews DH projects. Author of *Manuscript Fiction in the Age of Print, 1750-1900*.

- I enjoyed reading your chapters! (*More DH*) — useful for all *lit scholars*.
- Come a long way in your thinking on this project from your initial thesis proposal.
- I think that the work you're doing here is important and interesting. I'm certainly learning a lot about the databases you're working with, and I appreciate the fact that in Chapter 2 you manage to write about them in a way that appeals even to a non-DH scholar like myself.

GENERAL COMMENTS:

1. TITLE

- Was: *The Digital Archive and Print Politics, 1789-99*
- Now: *Canonical Corpora of English Literature, 1789-99* → falls flat / doesn't pique my interest

Also, "Canonical" → a problem? [Will come back to this.] [Corpora are neither canonical nor are they perceived as such.]

The word "Politics" could still be very useful for you. → The Politics of Digital Corpora, 1789-99

- This decade immediately signals French Revolution for c18 and Romantic scholars = Need to address
- Could use "politics" and "revolution" creatively to discuss how you're using the decade to examine a different kind of politics (re: database creation/construction) and a different kind of "revolution" (re: digital/research).

CHAPTER 1 COMMENTS:

Pg. 1: 1st para. = Still not sure what "popular" means.

Rec: Sentence one – *If one were to define popularity in terms of X, then according to the ETSC, the most popular English authors of . . .*

Pg. 2: "Normal" = what do you mean by this? Need to tease out.

Pg. 2: You write that "corpus-building has become the new canon-building" Has it? Do scholars think this? Are these texts treated as "ideal" representatives?

→ For me, reference only. Always approach them with a degree of skepticism. –Q. the quality of the search function, the text itself.

→ Is this not true for other scholars?

I think you're absolutely right that all literary research is now mediated by search algorithms and databases, but I wonder about the varying degrees to which it is directing that research, and I wonder if it might be useful to investigate that and spell it out more here.

Pg. 2: You write, "I focus specifically on writing printed in England between 1789 and 1799." But don't state why.

Pg. 4: For me, there was an odd jump between the first and second paras. insofar as you introduce Smith, and then move swiftly to the Gothic. Too much too fast. There are a couple of ways to mitigate this:

- You write that Smith has a long history of contentious reception, and refer to issues of "seriousness, popularity, and women's writing." But this for me is a bit vague. What if you were to say something like Smith, with her connection to the Gothic, to literature of sensibility, and to women's writing—genres that in the c19 struggled to achieve / retain legitimacy . . .
- Pg. 4: Because you're sectioning up your chapters, I thought it would be helpful to begin a new section with the para. "The problem of evaluating literature" . I.I From Canon to Corpus.

Begin with the idea of canon formation and print anthologies.

Move to the ways in which those canons have been / were shaped by reviewing culture.

Pg. 5: This odd jump continues onto pg. 5 when you then introduce the theatre and the French Revolution by way of Taylor.

Again, too much too fast.

Also, book reviewing and reviewing for the theatre = different. You're right that they're both part of a reviewing culture, but they are different subcultures. Can you find another source for a literary reviewing culture? Like Leigh Hunt and *The Examiner*? --Jonathan Mulrooney

p. 16: "through which I attempt to understand the 1790s" = What about the 1790s? ~~to see~~

SECTION 1.4 SCOPE

p. 21: Paine → not *Common Sense* but *Rights of Man* – Also Burke, Godwin, Wollstonecraft (Joseph Johnson Circle = publisher Joseph Johnson)

Then we switch to WW and Dickens! ~~Delete~~

~~Instead,~~ Need to spend some time on the French Revolution and politics and its influence on publication.

Then can move to the idea that this was also an important literary decade – WW, Mary Robinson, Charlotte Smith, Ann Radcliffe, Helen Maria Williams -- decade of the woman writer.

→ See Angela Keane's *Women Writers and the English Nation in the 1790s* (Cambridge, 2000) = addresses the imptc. of these women in this decade with reference to issues of canonization. She also has a book called *Revolutionary Women Writers: Charlotte Smith & Helen Maria Williams*

I think you can create a bridge here in this section between the French Revolution and the flourishing of women writers, many of whom wrote directly about the Revolution, including Charlotte Smith.

But even then, you need to make clearer your answer to the question, Why Charlotte Smith?

Prolific in 1790s – p. 23 (but it was also v. productive for other women writers)

French Revolution

- Poem *The Emigrants* (1793) – French Revolution
- Novel *Desmond* (1792) – French Revolution

Questions of Genre (p.30)

Tease Out / Make clearer?

Also, you suggest that an urgent concern for you is that women writers like Smith are underrepresented in the databases. I think, then, it's important that you take on a comparative approach. Do the databases capture the work of male authors in this period better? (will you do this in CH 3?)

Lyrical Ballads might not be your best bet, since it was produced in 1798?

p. 22: 1st sentence 1st full para. = Quite a claim! "The winners of the cultural capital game of the Romantics in poetry and Walter Scott in prose."

Romantics = broad category. Do you mean the Big 6?

Walter Scott = prose

What about Austen?

Also, this has changed.

Canonical Corpora of English Literature, 1789-99

Lawrence Evalyn
May 6, 2020

Chapter One

1.1. Introduction

According to the English Short Title Catalogue (ESTC), the most popular English authors of the 1790s were Thomas Paine, Hannah More, John Wesley, and William Shakespeare. Of course, this claim immediately falls apart on further scrutiny. In fact, by the metric of 'unique entries in the ESTC database,' the most popular author of the decade is by far Great Britain, followed by Great Britain, Great Britain, Great Britain, and King George III.¹ Paine, More, Wesley and Shakespeare are only able to rise to our notice if we intervene in the dataset to filter out all authors whose names contain the phrase "Great Britain"; otherwise, Shakespeare is outnumbered by the House of Lords and by the Church of England. And a single paragraph cannot contain all of the reasons that the quantity of unique entries in a database would not correlate with any useful definition of popularity -- although later parts of this dissertation will undertake to enumerate them at greater length. These claims demonstrate that a poorly formed question will produce a useless and stupid answer even (or perhaps especially) if computation is used to

¹ More specifically, these "authors" are "Great Britain, Parliament," "Great Britain," "Great Britain, Parliament, House of Commons," "Great Britain, Lords Commissioners of Appeals in Prize Causes," and King George III. After King George comes Thomas Paine and Hannah More, and then it's "Great Britain, Parliament, House of Lords" and "Church of England."

answer it. This dissertation is dedicated to the formulation of better questions. I am interested in the limits of the generalizations that we make, both in “distant reading” research and in non-digital scholarship. I take as my starting point the contention that, in order to identify what is “popular” or “important,” we must also understand what is normal. At its core, my question is: given that it is not possible to read everything (or even most things), how do we, and how *should* we, determine what to read, preserve, study, and teach? This “question” is, of course, many questions: what we do is by no means what we *should* do; what we read is not necessarily what we study or teach. *The question of what to read, preserve, study and teach* It is also an old, nearly an old-fashioned question. The current moment of self-reflection in the field of Digital Humanities, however, provides a timely reason to revisit it. Even literary scholars who do not carry out “Digital Humanities” research are impacted by the corpus-building choices of major digital resources, since all literary research is now mediated at some level by search algorithms and databases, even if this mediation is as small as looking up the holding libraries for physical copies of texts. It is therefore relevant to the field as a whole if, as I contend, corpus-building has become *a form of* the new canon-building: an invisible and naturalized process of selecting texts for idiosyncratic and historically-specific reasons, and then treating those individual texts as ideal representatives of an imagined “whole” of literature. *Has it? Do scholars do this?*

Despite the crucial importance of corpus-building to the interpretation of “distant reading” research, it is often extremely difficult to know what is in a corpus. Even large institutional resources used by many scholars provide little context for their choices of what to include or exclude. These hidden choices are particularly problematic when historical selection factors might have led to the creation of databases which re-create social inequalities. I focus specifically on writing printed in England between 1789 and 1799, to explore how works from this eleven-year “decade” have been selected as important, literary, or popular. For this period, *Why?*

(~~ESTC~~) ^{on}
the English Short Title Catalogue provides basic bibliographic data for nearly 52,000 titles, but
the Eighteenth Century Collections Online Text Creation Partnership ^{ok} ~~(ECCO-TCP)~~ corpus of XML-encoded
full texts includes fewer than 500 titles. This difference raises the question: why were the other
51,500 titles *not* considered worth the investment of scholarly effort? And with particular ^{urgency?} / ^{urgency?}
urgency: do the most invested-in resources underrepresent women? ^{written} My experiments examine six
major databases to answer these questions: The English Short Title Catalogue (ESTC),
Eighteenth Century Collections Online (ECCO), the Eighteenth Century Collections Online Text
Creation Partnership (ECCO-TCP), Google Books, Project Gutenberg,² and HathiTrust. For each
database, I download their holdings identified as printed in England 1789-1799.³ I identify how
many titles the database attributes to each year. I calculate how many works are attributed to
male, female, or unknown authors.⁴ These very simple pieces of information, when they differ
widely between databases,⁵ provides the basis for an initial analysis of the assumptions and
limitations of each database. I then examine the contents of each database more closely, to
compare the inclusion of broad categories of writing like poetry, drama, prose fiction, and
ephemera.⁶ Identifying these categories of writing within each corpus reveals a predictable
preference for "literary" forms such as novels and poetry [in the smaller databases.] This
preference for particular kinds of writing might explain changes in gender representation of

² Google Books and Project Gutenberg are not, of course, traditionally "scholarly" resources, but that is why they form an informative contrast with the other resources examined.

³ This calculation is carried out by manually examining the metadata of the six corpora I have acquired.

⁴ This calculation is carried out by a small, simple program I am writing, described in Appendix A. Because the program just simplifies a straightforward process of counting, it is only lightly theorized in the dissertation itself.

⁵ I already know that 'titles per year' are distributed fascinatingly differently between ECCO, ECCO-TCP, ESTC, and HathiTrust

⁶ This calculation is carried out by a larger, more complex program I am writing, applying topic modelling to the titles of works. Because it makes several major interpretive choices, it is theorized and discussed in detail when it is applied.

smaller databases. If the novel is the domain of women, for example, a corpus can underrepresent women by under-representing novels. Or it could include a representative number of novels, but disproportionately include novels by men. My investigation allows me to identify the patterns of selection. To ground my analysis in specifics, I take Charlotte Smith as a case study author. Smith has a long history of contentious reception, rooted in debates about seriousness, popularity, and women's writing. I revisit them to see how her career and reception might be interpreted through a new lens. I do so, in part, to challenge the contrast drawn between 'popular' and 'serious' writing, especially in the historical evaluation of women's writing as literary.

legitimacy?

w/ her connection to the gothic is literature of sensibility and?

Why so? (need short explanation)

critical

Maybe mention Smith as affiliated w/ gothic and other motifs eg sensibility so that when we get to the gothic in the next of its keys of a jump.

NEW SECTION T.I FROM CANON TO CORPUS - BGS' w/ IDEA OF CANON

The problem of evaluating literature is not a new or a simple one. In the eighteenth century, the debate took the form of urgently needing to distinguish 'trash' from 'treasure'. Michael Gamer, in *Romanticism and the Gothic: Genre, Reception, and Canon Formation*, highlights the role of the eighteenth-century reviewer as a crucial mediator between the writers and readers of books. Importantly, although the assessments take the form of reviews of individual works, Gamer also argues that the critics' objections are in fact "a regulatory discourse – carried out under the fiction of paternalistic advice to a given gothic writer, but functioning as an implicit threat to other readers and writers" that affiliation with the gothic comes with "cultural costs" (42). The gothic stands in as a proxy for any kind of "popular" reading that takes place "in the absence of formal education and training" (57), so a denunciation of a gothic work becomes a reaffirmation of class-based literary hierarchies. In other words, these reviews create and affirm the cultural capital of a category of 'serious' literature. Gamer is only concerned with the gothic and romanticism, but the overall regulatory function of literary reviewers as moral arbiters—and the stock conventionality of their objections, which do not affect the actual production or

How has that Canon evidenced itself historically? Anthologies now receiving then

Also brings ↑ French Revolution
Φ Theatre problematic
↓ Another source for literary reviewing culture?
consumption of the works attacked— applies to most forms of writing in the period. For example, George Taylor sees the same dynamic in the theatre. In *The French Revolution and the London Stage*, he argues that, “[c]ritics might make sharp comparisons” between the many kinds of entertainments that were staged, “but little of the programme was dismissed [by audiences] as ‘trash’, or ‘immoral’, or irrelevant ‘fancy’” (3). Taylor sees the repetitive discourse of eighteenth-century literary critics as proof of a larger social divide: “Disagreement as to what is trash and what is treasure suggests cultural crisis, when values are put under question by social stress or political conflict” (3). Gamer and Taylor both suggest that moral judgment of literature by its critics was driven by social friction, rather than by the aesthetic distinctions which they claimed as their motivation. *vague*
What do you mean by this?

Jonathan Mulrooney book Romanticism and Theatre Experience
In other words, Gamer and Taylor both affirm the key conclusion of John Guillory’s *Cultural Capital: The Problem of Literary Canon Formation*, that “in fact ‘aesthetic value’ is nothing more or other than cultural capital” (332). *quotes* Guillory’s sociological history of literary canons is a well established part of literary studies, which will take on new dimensions as I apply to to the current moment of digital databases. In the eighteenth century, he argues, the cultural capital of vernacular English literature is defined by its use within the school system to enable and restrict social mobility. English vernacular literature first begins to accumulate cultural capital in middle-class schools where it is “a substitute for the study of Greek and Latin, but with the same object of producing a linguistic sign of social distinction” (97) that would allow readers to improve and signify their social standing. The public re-assessment of literature described by Gamer and Taylor is, for Guillory, “the first crisis in the status of the vernacular canon, the problem of assimilating new vernacular genres such as the novel” (xi), which seem in danger of

affording too much social mobility by offering too little literary distinction for social elites.⁷ The 'solution' is institutionalization, in which "the school becomes the exclusive agent for the dissemination of High Canonical works," and therefore, he argues, "the prestige of literary works as cultural capital is assessed according to the limit of their dissemination, their relative exclusivity" (133). Under this system, 'serious' literature may not be identifiable linguistically, but it can still be identifiable by the difficulty of accessing it. This history of canonization has important implications for the field of literary study. As Guillory himself insists, if the aesthetic value of a text is determined by the social operations of class, it undermines the notion of literature itself as a category of writing distinguishable in aesthetic terms from non-literary writing. Guillory's book is motivated by the canon debates of the 1990s, which were driven by an urgent re-valuation of literature by women and people of colour.⁸ His response insists that it is untenable to conceive of the problem in terms of increasing the 'representation' of individual works or authors within existing systems. Instead, for Guillory problem lies in the institutionalization of literature itself. "If literary criticism is ever to conceptualize a new disciplinary domain," he says, embedding his prescription in that "if," "it will have to undertake first a much more thorough reflection on the historical category of literature; otherwise I suggest that new critical movements will continue to register their agendas symptomatically, by ritually overthrowing a continually resurgent literariness and literary canon" (265). In other words, assigning the cultural capital of "literature" to different works cannot change the underlying system.

⁷ 'Too much' and 'too little' are here, of course, defined from the point of those with cultural capital which they wish to maintain.

⁸ Part of Guillory's argument is that, although the rhetoric of the canon debates generally sought to re-value authors of any number of oppressed categories, often using the phrase "gender, race, and class" as a single unit, the work undertaken was in fact unable to address class, since class operates differently from gender and race.

Perhaps indicating that Guillory was correct, twenty years later, we are still debating the need for “literary criticism ... to conceptualize a new disciplinary domain” (Guillory 265), now in the context of computation. The reconceptualization of literary study itself is at the core of Franco Moretti’s coinage of ‘distant reading’: the problem for which “[r]eading ‘more’ seems hardly to be the solution” (“Conjectures” 55) is the problem of conceiving of *world* literature, rather than the “canonical fraction, which is not even one per cent of published literature” (55). His new methods are meant to enable literary studies to examine a new object. The field of distant reading has been moving away from Moretti himself. However, it is still shaped by the attempt to redefine the disciplinary domain of literary studies. In many cases, the new domain is no longer the “canon” but the “corpus,” a collection of texts which are studied *en masse* for macroanalytical insights. Katherine Bode, for example, in “The Equivalence of ‘Close’ and ‘Distant’ Reading,” argues that Moretti and Matthew Jockers replicate the approaches of New Criticism with their corpora, and calls for “a new scholarly object of analysis” (79) that directly examines historical and textual context of corpora as representations of “literary systems” (97). Lauren Klein, too, treats the textual corpus as the new object of literary analysis requiring curation, contextualization, and interpretation. Her critique argues that “it’s not a *coincidence* that distant reading does not deal well with gender, or with sexuality, or with race,” but also that these failings are not inevitable: “it’s not that distant reading *can’t* do this work,” she insists, “it’s that it’s yet to sufficiently do so” (n. pag.). Bode, too, despite her strong critique of distant reading as it has been practiced by Moretti and Jockers, does not blame distant reading itself. Distant readers like Moretti and Jockers, she argues, “while claiming direct and objective access to ‘everything,’ ... represent and explore only a very limited proportion of the literary system, and do so in an abstract and ahistorical way” (78). Klein, like Bode, calls for “more corpora—

more accessible corpora—that perform the work of recovery or resistance” to allow research “beyond quote ‘representative’ samples, which tend to reproduce the same inequities of representation that affect our cultural record as a whole” (n. pag.). This framing re-creates, at the site of the corpus, the identical narratives of exclusion and representation which were previously located in critiques of the canon.

The relocation of the debate from the canon to the corpus is not without grounds. As this dissertation will explore in depth, challenges to the technological accessibility of texts have created new hierarchies, and a new “great unread.” Each archive represents a unique set of choices in response to the same sets of questions: what to include, why, how; what to make accessible, why, how, to whom; what, in the end, makes a text matter, and what we are meant to *do* with texts. For example, the English Short Title Catalogue records 51,965 titles printed in England between 1789 and 1799. The corpus most commonly used for DH work on eighteenth-century literature, ECCO-TCP, includes only 466 titles for that same time period. What are the other 51,499 titles, why are they accessible in the ways they are, and what does it mean for digital eighteenth-century studies that they are not included? Although the examination of databases prompts similar hypotheses of exclusion as in longstanding conversations about canons, digital databases do not simply replicate new canons. [By the end of my dissertation, I will be able to state here what IS happening — something structured by related logics of access and prestige, and related simplifications of historical complexity, and related *institutional* replication of privileged texts.. But very importantly different, too, since we don’t *read* databases.] In a series of computational and non-computational research processes, I examine six databases of eighteenth-century texts to learn about four eighteenth-century authors, and I examine four eighteenth-century authors to learn about eighteenth-century databases. This

> This has changed ?

dissertation, therefore, takes place within three scholarly conversations: the digital humanities, as an increasingly self-reflective set of practices; eighteenth-century studies, and the challenges presented by the 1790s; and the frameworks of reparative reading within queer theory which seem to offer valuable resources for both. The remainder of this chapter will describe in more detail the relevant scholarship shaping my frameworks, and then introduce my chapters by introducing my four case study authors.

1.2. Frameworks

My work takes a critical algorithm studies approach to digital databases of eighteenth-century literature, examining the structural assumptions of the most-used resources (including some that scholars don't like to admit to using). I close read the database structures, file formats, and historical documentation for the English Short Title Catalogue, Eighteenth Century Collections Online, the Text Creation Partnership, HathiTrust, Project Gutenberg, and Google Books, to examine how each resource's algorithmic definition of a "book" (and the information that might matter about a book) is shaped by the material, historical conditions of each organization's development. My initial research question was, by Eve Kosofsky Sedgwick's definition, a classically paranoid approach: I sought to expose the under-representation of women's writing underlying apparently "neutral" digital infrastructures. This question carried the combined urgency and futility of paranoid critique: urgent, because an unfair database would expose an unfair society; and futile, since the research could only be motivated by the conviction that its answer was already known. My paper will touch briefly on some specifics of this research and my findings, as the basis for a broader discussion of critical algorithm studies, and the project of imagining reparative algorithm studies.

One of the current problems of critical algorithm studies is how difficult it is to move from

critique to action: it seems that no matter how carefully we dissect the flaws of oppressive computational systems, we cannot opt out of them. Excellent work by scholars like Wendy Hui Kyong Chun and Safiya Noble, for example, meticulously historicizes computational systems, and there is real value to the denaturalization of the systems they thus reveal. But this work relies on the paranoid logic of exposure, and I am interested in other attitudes. In my examination of digital infrastructures for eighteenth-century studies, I take a brief detour through Marxist thinking (via Bourdieu and John Guillory) to diagnose a deep tension between capitalist and anticapitalist value systems as the likely cause of the flaws in these systems today. I then am to move beyond the obvious paranoid critiques prompted by this observation. I confess that, at this stage, this is the point at which my thinking remains speculative— but I feel certain that the right direction lies in queer strategies of creative reappropriation, subversion, and resistance.

The theoretical frameworks of this dissertation are drawn from the fields of feminist DH and queer DH, and from non-DH schools of thought which seem to offer valuable tools. My core motivating framework, as I conceptualize my work, is that of reparative reading. ~~Eve~~ Sedgwick's "Paranoid Reading and Reparative Reading" persuasively describes in the dominance of paranoia in literary criticism, and attempts to sketch an alternative in what she terms reparative reading. A paranoid rhetoric of exposure and critique strikes me as the most obvious narrative to structure this dissertation's investigation of the uneven institutional valuation of different writing. ^{present in digital corpora} However, these obvious critiques also require rejecting many generations of sincere work by ~~my~~ ^{n.} fellow academics, without necessarily offering new discoveries of value to replace them. One experiment of this project, not yet complete, is to articulate an assessment of the limitations of contemporary digital resources which nonetheless allows ~~those~~ ^{for} those resources to be recuperated. My touchstones are two descriptions from Sedgwick's original chapter:

You already mention her on previous page
No need for 1st name

The desire of a reparative impulse... is additive and accretive. Its fear, a realistic one, is that the culture surrounding it is inadequate or inimical to its nurture; it wants to assemble and confer plenitude to an object that will then have resources to offer to an inchoate self. (149)

What we can best learn from such practices are, perhaps, the many ways selves and communities succeed in extracting sustenance from the objects of a culture - even of a culture whose avowed desire has often been not to sustain them. (150-151)

What Sedgwick describes, here, is a “desire,” not a methodology. I therefore understand “reparative reading” to refer, not to a precise set of practices, but to a position one might occupy in relation to a text. What I posit is also a desire: that my methods here can provide useful practices for others. The reparative position is a generous one, both in terms of giving of oneself to a text, and in terms of seeking a text’s strengths over its weaknesses. What I learn from Sedgwick, therefore, that *attention* is the first step toward *caring*, and that non-judgment can be more informative than rejection.

I have mentioned moving away from critique as well as from paranoia: in rethinking the role of critique, I draw upon the work of Rita Felski, and the theories of “surface reading” described by Sharon Marcus, Stephen Best, and Heather Love. Felski, in her article “After Suspicion” and then further in her monograph *The Limits of Critique*, seeks to attend seriously to literary attachments, including our own attachments as critics. Felski’s approach to these attachments is essentially sociological, drawing heavily on Bruno Latour’s actor-network-theory, and thus involves almost no close reading. “Surface reading” positions itself as an alternative to “symptomatic reading”; rather than seeking to expose hidden truths concealed within texts, it attempts accurate descriptions that “make visible what is invisible only because it’s too much on the surface of things” (Best 13). The analogues to reparative and paranoid reading are obvious, but not perfect: all paranoid reading is symptomatic, but not all symptomatic reading is paranoid. Reparative reading, as described by Sedgwick, is often still interested in ‘deep’ meanings of

texts, in which striking textual features can be interpreted to locate additional meanings. Felski's readings are often symptomatic in this way. In contrast, "surface reading," as Heather Love describes, pursues "a turn away from the singularity and richness of individual texts" (374), seeking descriptions that are "complex and variegated, but not rich, warm, or deep" (378). Love's disavowal of "richness" here is part of her attempt to move away from "the ethical charisma of the literary translator or messenger" (374) who characterizes the paranoid, critical figure that both Sedgwick and Felski also seek to escape.

Love's later article, "Close Reading and Thin Description," provides a more precise articulation of the kind of close reading that she calls for, in which an "exhaustive, fine-grained attention to phenomena" (404) enables "taking up the position of the device; by turning oneself into a camera, one could—at least ideally—pay equal attention to every aspect of a scene that is available to the senses and record it faithfully" (407). Although Love is uninterested in "distant reading" as synonymous with Moretti (Love 411), this invocation of the mechanical implies, I argue, an obvious potential for computation. The actual *practice* of computational research requires a great deal of laborious, intimate encoding. The researcher must occupy a "mechanical" position of receiving inputs and responding to them consistently over time, whether entering details in a spreadsheet with a consistent taxonomy or running the same program over multiple datasets.⁹ Love says:

Good descriptions are in a sense rich, but not because they truck with imponderables like human experience or human nature. They are close, but they are not deep; rather than adding anything 'extra' to the description, they account for the real variety that is already there. (377)

⁹ Appendix B ("Methodology") contains many examples of these algorithmic procedures executed by the human researcher and the computational programs in concert. The act of writing a program is an iterative process of delegation.

A computational model is unlikely to “truck with imponderables,” but it *absolutely must* “account for the real variety that is already there” or else the code will simply fail to run. If you are forced to manually encode your assumptions into a system, you are forced to confront what they are. Even deleting or ignoring information is still a way of “accounting for” it in the coding process: some part of the program will have to say, in effect, ‘if I get an input that doesn’t match what I expect, discard it.’ Choosing to ignore contradictory or difficult information carries the assumption that this information does not ‘count,’ or does not matter to the question at hand. The choice faced by scholars is how to address our encoded assumptions. The encounter with variety does not in itself produce nuanced results: it is possible to selectively ignore any uncomfortable details. But it is also possible to do computation reflectively, asking not “how can I make this work the way I want?” but “where do my assumptions encounter resistance?” and turning one’s attention to the nature of the resistance. Integrating this reflection into the research process can allow a scholar to avoid both the pitfalls of “conquering” their material and of claiming an algorithmic grasp of “objective” truth.

SP^{ing}.

nice

To bring these principles into the field of Digital Humanities by way of an example, I want to offer an alternative genealogy for the practice of distant reading itself. Rachel Buurma and Laura Heffernen provide a valuable history of Josephine Miles as the first ‘distant reader’. Miles’ history, briefly, is as follows:

fellowship?

In the 1930s, as a graduate student at Berkeley, she completed her first distant reading project: an analysis of the adjectives favored by Romantic poets. In the 1940s, with the aid of a Guggenheim, she expanded this work into a large-scale study of the phrasal forms of the poetry of the 1640s, 1740s, and 1840s. In all of this distant reading work, Miles created her tabulations by hand, with pen and graph paper. She also directed possibly the first literary concordance to use machine methods. In the early 1950s, Miles became project director of an abandoned index-card-based Concordance to the Poetical Works of John Dryden. Partnering with the Electrical Engineering department at

Berkeley, and contracting with their computer lab and its IBM tabulation machine, Miles used machine methods to complete the concordance. It was published in 1957, six years after she and several woman graduate students and woman punch-card operators began the work. It was thus begun around the time that Busa circulated early proof-of-concept drafts of his concordance to the complete works of St. Thomas Aquinas, and published 17 years before the first volumes of the 56-volume *Index Thomasticus* began to appear. (Buurma and Heffernan)


Buurma and Heffernan bring Miles' history to our attention not simply because Miles predates Roberto Busa, whose *Index Thomisticus* is often credited as the first large scale computational literary study.¹⁰ Rather, they emphasize, Miles' origin story for computational literary study "can stand as an example of how we might write a history of literary scholarship that does not center originality and individual accomplishment" (n. pag.). Unlike Busa, Miles not only gave authorship to the (female) graduate students who carried out much of the labour of creating the concordances, she also thanked and credited the (female) punch card operators who encoded the resulting data.¹¹ Moreover, when talking of Penny Gee, one of the female staff members of the computer lab, Miles praises her as "'very smart and good' and—most importantly—a true collaborator, as opposed to those 'IBM people from San Jose' ... 'I've never been able to connect with them,' Miles explains, 'though I did with Penny Gee. She really taught me'" (n. pag.). Of the positive qualities highlighted here, only one, "smart," is traditionally valorized among literary critics: to be "good," a "collaborator," who can "connect" and "teach" — these qualities are often seen as irrelevant to the singular authority of the figure of the critic, but they

¹⁰ Indeed, Buurma notes, "There are good reasons, of course, that scholars and journalists like to begin with Busa: he was the first concordance-maker to automate all five stages of the process, in 1951," and he intentionally foregrounded and publicized the innovative nature of his work. \cite{Buurma:2018wt}

¹¹ In the interest of preserving this history of citation, the students were Mary Jackman and Helen S. Agoa, credited on the cover of the published Dryden index. (Miles herself attached her name only to the preface.) From the computer lab staff, Miles particularly thanked Shirley Rice, Odette Carothers, and Penny Gee.

are core to a reparative practice. Miles' work, too, struggled to find appreciation "among literary critics who viewed her datasets as merely preparatory to the true work of evaluation" (n. pag.).

What's crucial, to use computational reading reparatively, is to use it *reflectively*. The desirable kinds of computation which I describe above will not happen inevitably. Here I draw upon the rich body of work emerging in critical algorithm studies, which examines (and attempts to reform) the human elements of computational algorithms. Any methodology is, to a certain extent, an "algorithm," in the loose definition of 'a series of pre-defined steps to be carried out'. But computational algorithms differ from "algorithms" implemented by humans. Computational algorithms have two key vulnerabilities: first, their operations are less easily scrutinized; second, their results are more easily trusted. The second vulnerability — the cultural aura of empirical trustworthiness which accrues to anything 'computational' — is another flavour of the same vulnerability that Drucker describes with 'data' generally. Because the human agents who designed and trained any given algorithm appear to be absent from its operation, the algorithm appears able to discover truth directly. This is how Daily Wire reporter Ryan Saavedra was able to tweet with disdain that "Rep. Alexandria Ocasio-Cortez (D-NY) claims that algorithms, which are driven by math, are racist" (@RealSaavedra): anything "driven by math," he assumes, must be incapable of human fallibilities like racism. But as Safiya Noble shows extensively in *Algorithms of Oppression*, algorithms by default reproduce, and can easily exaggerate, the assumptions and biases of the culture in which they are made (CITE). In other words, in a racist world, algorithms *are* racist — and sexist, and duplicative of all other systemic inequities. To analyze an algorithm, one must articulate the implicit argument underlying the assumptions that allow it to operate — what Ian Bogost would call its procedural rhetoric. As Katherine Bode's recent work on data-rich literary history has shown, digital infrastructures themselves contain an



implicit procedural rhetoric, even an argument, which must be addressed.

Critical algorithm studies is therefore a crucial background for my work — but “critical” is literally in the name of the field, and I still seek to be post-critical and reparative. As I encounter the limitations of the various information and tools through which I attempt to understand the 1790s, my goal is to do something other than facilely observe that they are limited. Instead, I want to identify the best ways to continue building on their foundations. In a digital humanities context, a focus on building connections can be mundanely practical: typing indexes from print works into spreadsheets, correcting errors within datasets, writing programs to process metadata: all of these maintain the functional usability of existing resources in new contexts. When this kind of extended, detail-oriented labour is combined with serious reflection on the histories and possible futures of these resources, I contend, they bring us to new knowledge. In this, maintaining and using digital resources is also a way to repair them — and to produce reparative readings of their contents.

1.3. Methods

This dissertation undertakes computational distant reading. At every possible point, however, the underlying methodology will be made visible, and its assumptions scrutinized. The bibliographic histories of my multiple corpora are explicit objects of inquiry. Much of the code underlying this project I have written myself. Some has been written at my request. In every case where the code is available to me, the program itself appears in Appendix A (“Codebase”), accompanied by a plain language explanation of how it operates. Where I have used closed-source software, Appendix A contains an explanation of my best guess at its underlying process. My exact use of these tools — sufficient for another to replicate my work — is provided in Appendix B (“Methodology”). These details are explicated in full in the appendices in order not

to over-burden the body of the dissertation, but they are by no means *confined* to the appendices. Computation is not a “black box” to be consulted for simple answers, but is inextricable from my reasoning and argument.

My attention to the *sources* of digital knowledge creation comes, in part, from Johanna Drucker, and her distinction between “data” and “capta.” Drucker, in “Humanities Approaches to Graphical Display,” specifically addresses the digital humanities practice of creating, and then close reading, data visualizations. She argues that the tools for visual representation which may be effective in the sciences cannot be simply and uncritically transposed to humanistic subject matter. When an experiment is presented as a ‘data visualization,’ she says, “the rendering of statistical information into graphical form gives it a simplicity and legibility that hides every aspect of the original interpretative framework” (8). In fields where the readers of such charts are also frequent creators of charts, and where norms exist to explicitly describe one’s interpretive frameworks in a methodology section, the simplicity and legibility of an individual chart may be a benefit which does not impede complex scrutiny of the information it presents.¹² In a field like literature, however, the “graphical force” of something like a network graph or even a simple pie chart “conceals what the statistician knows very well — that no ‘data’ preexist their parameterization” (8). Drucker problematizes the term “data,” the etymology of which presents it as a “given” which is stable and independent of observation. She proposes that humanities visualizations embrace, instead, the framework of “capta,” that which is “‘taken’ actively” (3), “fundamentally codependent, constituted relationally, between observer and observed phenomena” (50). Drucker’s assessment shapes my own prioritization of qualitative and

split
inf.

¹² It may also be the case, of course, that even fields with a long history of graphical display would benefit from greater scrutiny of the evidence they use; see: the Data Dinosaur. But this is beyond the remit of what an English PhD can address.

reflective computational research. The term “capta” itself has not seen uptake in subsequent digital humanities scholarship, even in cases where scholars explicitly take Drucker’s warnings to heart. Accordingly, for clarity, this dissertation will continue to use the more usual term “data” to refer to the information gathered for analysis here. However, as I integrate and compare a wide variety of data from many disparate sources, a preliminary task of my analysis is always to determine, as precisely as possible, how the information was captured and quantified. ✓

Additionally, all of the figures presented in this dissertation are of my own design. My design praxis is informed by the work of Edward Tufte and Alberto Cairo, both of whom provide practical design advice in service of demystifying the visual rhetoric by which graphs present their arguments.¹³ Neither Tufte nor Cairo is a scholar of media studies; rather, they are professional practitioners of ‘data visualization’ who reflect critically on the assumptions of their work. Tufte’s work primarily strives to correct badly designed data visualizations, and the dangerous decisions that bad design can lead people to. His most famous example is an analysis of the engineers’ report at NASA which led to the ill-fated launch of the Challenger space shuttle in 1986: as his extensive visual analysis argues, the engineers (untrained in graphic design) unintentionally obfuscated crucial information about the day’s launch conditions. The poorly designed graphics these engineers produced made the launch appear low risk to their superiors; despite the engineers’ strong warnings, their verbal argument was disregarded in favor of their accidental graphical argument. As Tufte demonstrates, a few simple alterations of their graphic design would have made it obvious that the day’s unprecedentedly low weather was extremely

¹³ I cite Tufte and Cairo as the thinkers whose design philosophies best accord with my own current understanding of the work and craft of persuasive data visualization, but my actual practical training as a graphic designer is indebted to Judith Galas, Sonia Davis Gutiérrez, and Tom Hapgood.

dangerous, and potentially averted disaster \cite{Tufte:2001vw}.¹⁴ Tufte's six principles of design¹⁵ primarily seek to guide undertrained designers away from misleading themselves. Cairo, following on Tufte's work from the perspective of an active journalist, more often turns his attention to successful designs which mislead their audiences intentionally. His forthcoming book, *How Charts Lie*, addresses the readers of infographics with insights into visual literacy \cite{Cairo:iklksuMr}. His preceding book, *The Truthful Art*, addresses the creators of good faith infographics with insights into visual manipulation \cite{Cairo:2016uv}. Cairo draws a distinction between "data visualization" and "infographics": "an infographic tells the stories that its designer wants to explain, but a data visualization lets people build their own insights based on the evidence provided," summarized more succinctly as "infographics to explain, data visualizations to explore" \cite{Cairo:2014tl}. Using this terminology, my argument will proceed with infographics in the body of the dissertation as curated figures to support my argument, with fuller data visualizations available in Appendix C ("Data") to allow further exploration. Following in both Tufte and Cairo's footsteps, I conceive of the figures throughout this dissertation as rhetorical devices. In service of arguing honestly, therefore, my designs — in the body of the dissertation and in Appendix C — are accompanied by footnoted explanations of my design rationale.

This dissertation understands archives, bibliographies, anthologies, and corpora to all be,

¹⁴ Tufte is careful not to blame the engineers for being better at engineering and systems analysis than they were at design: rather, this example shows that design is a skill that involves expertise; when designs matter, people with that expertise need to be involved.

¹⁵ 1. show comparisons, contrasts, differences 2. show causality, mechanism, explanation, systemic structure (intervention relies on manipulable causality -- can't do anything with the information without causality) 3. show multiple variables (3 or more) -- the world is multivariate 4. *completely integrate* words, numbers, maps, graphics, etc, etc. Provide information at exact point of need 5. documentation must thoroughly describe evidence and its sources, provide complete measurement scales 6. presentations succeed based on their content. for better presentations, get better content.

variously, *models* of an imagined object of study. In the language of social science, these models might be described as ‘samples,’ which are intended to permit discoveries about an underlying ‘population’ by being ‘representative’ of that population’s features. Only the language and not the methods of social science need to be imported here, since it has long been ordinary practice in literary studies to select and examine representative texts for insights about larger movements¹⁶. A work like Ann Tracy’s bibliography *The Gothic Novel 1790-1830*, for example, clearly names the population of works which are of interest to her: all Gothic novels published between 1790 and 1830. She tentatively defines her principles of selection as _____. But in providing detailed information on 208 texts — mostly Gothic, mostly novels, mostly between 1790 and 1830 — Tracy obviously does not claim to have presented all that might belong within this population. Instead, her book operates as a model of the underlying population, which can be queried for further insight into ‘the Gothic novel, 1790-1830’ only so long as one keeps the limits of the model in mind. Indeed, by presenting plot summaries and bibliographic data, rather than reproducing the novels in full, Tracy provides a model of a model. One challenge to studying these models is that they present a “moving target”: even a bibliography or anthology is subject to change through successive editions (not to mention their now-common digital supplements), and a digital database has the potential to change daily. I follow Kath Bode’s approach in *A World of Fiction*, in artificially “freezing” each resource for study, and presenting

¹⁶ Kath Bode and Leah Price have both described at length how textual editing and anthologizing, respectively, are literary methods of sampling. See: Bode, Katherine. *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press, 2018. And Price, Leah. *The Anthology and the Rise of the Novel: From Richardson to George Eliot*. Cambridge UP, 2000.

my analysis as a description of a snapshot in time.¹⁷ Importantly, a model is a tool for thinking, and not necessarily a truth claim in itself: creating a model is a way of saying, 'it might be helpful to think of X as Y,' not an assertion that X is equivalent to Y. Willard McCarty articulates this important feature of models by stressing that a model's value is determined not by its exact correspondence with the object it models — if it were possible to fully examine the underlying object, then no model would be necessary — but by the *fruitfulness* of its simplifications (CITE McCarty). Even a deeply incorrect model can be fruitful if its divergence from observed phenomena rules out an incorrect theory. As I examine the many existing models of 'English literature, 1789-1799,' and create several more of my own, I articulate the underlying assumptions of each model, and assess the fruitfulness of the results.

1.4. Scope

All of the computational work in this dissertation aims to identify, in as minute detail as possible, all works printed in England between January 1, 1789 and December 31, 1799. This eleven-year "decade" was a turbulent one across the Channel, encompassing the whole of the French Revolution, from the Estates General in 1789 to Napoleon's coup in 1799.¹⁸ In England, these events caused strong and variously nationalist reactions in a country which had so recently lost its colonies in America and feared that a French invasion could come at any moment. This is the decade of *Common Sense*, it is the decade of *Lyrical Ballads*; it is the decade of Hannah More, it is the decade of Ann Radcliffe; it was the age of wisdom, it was the age of foolishness; it

¹⁷ Because Bode is examining only one database, she is able to present a single date on which her data collection ceased. I have not been able to accomplish this, but for any given resource, will identify the date of that resource's "snapshot." This approach means that my observations may be out of date from the moment I make them, though one of my findings in chapter 2 is that many databases are currently changing more slowly than one might expect.

¹⁸ Although these events, of course, did not occur on January 1 or December 31, respectively, the entirety of 1789 and 1799 are both included in my study, out of sheer technological necessity.

1776!
Rights of Man (1791)

Thomas Paine's
Common Sense

W. Wordsworth's
Lyrical Ballads
vague

As Charles Dickens put it

beginning to end

William Godwin's Political Justice (1793)

? awkward

Edmund Burke's Reflections on the Revolution in France (1790)

was the epoch of belief, it was the epoch of incredulity. ^{His} Charles Dickens' now famous superlatives capture the tension often seen by scholars between 'Enlightenment' modes of writing and 'Romantic' or 'Gothic' modes, which are no longer neatly periodized as mutually exclusive.

Get rid of Dickens + this bit. Toss out this section

Scholarship on ^{late} eighteenth-century ^{literature} works often takes the form of evaluating or assigning the cultural capital of individual works, or, perhaps, analyzing the strategies by which they accrue or fail to accrue that capital. The winners of the cultural capital game are the Romantics in poetry and Walter Scott in prose. For example, Simon Bainbridge examines the decade and its poetry through the lens of war to identify "the attempts made by several writers to fill the role of national bard prior to Scott" (3). Both poetry and the poet, in his conception, are pursuing a particular kind of cultural capital that allows them to rise above their own popularity. Richard Cronin's *The Politics of Romantic Poetry* and Robert Miles, too, seem to treat Scott's intensely serious popular romances as the teleological end of the late eighteenth century birth development

Big 6?

see Gamers new book on conservatism

within the novel. These works follow a pattern established from the beginning with Kiely and ^{First name} Tompkins, of treating the novel as synonymous with the realist novel, and treating Romantic and especially Gothic novels as aberrations in the history of the novel, a problem which needs to be

First name

Romantic? unclear which

novel you referring to

explained away. E.J. Clery's *The Rise of Supernatural Fiction* has examined at length the historical conditions by which supernatural plot elements began to make limited claims to literary seriousness throughout the eighteenth century. The "rise" she describes is not an increase in volume and prominence of supernatural stories, since her starting point in 1762 (the Cock Lane ghost) is a major national phenomenon with many imitators. Rather, supernatural fiction 'rises' when it acquires cultural legitimacy. Michael Gamer has more recently expanded on how this 'rise' fuelled Romanticism's own rise. Gamer, like Bainbridge and Cronin, primarily

in what specifically? in supernatural fiction

Section needs fleshing out

all this? Smith, Robinson? Who has defined them as such?

examines Wordsworth and the 'winners' of the struggle for cultural capital: I, like Clery, am more interested in the 'losers.' Accordingly, although I take a highly successful literary writer as my primary touchstone, I also attend to much that is *not* literature, in order to better understand why it is not.

1.4.1. Charlotte Smith

To navigate the 1790s, I turn to an author whose career and works usefully focalize my core questions of genre, publics, and the status of literature: Charlotte Smith. Smith was highly productive in multiple genres throughout the 1790s, and had a complex and contested literary legacy after the 1790s. As literary scholars re-assess ideas about literary seriousness, popularity, and women's writing, our assessment of Smith has shifted as well. By examining their bibliographies with computational methods, I again ask how she might continue to look different if we look at her a different way. I particularly examine the extent to which digital resources have kept up with the re-evaluation of Smith as a central figure in British Romanticism.

[All of this introduction to Smith remains rough to give a gist of the kind of material I want to cover here; I will spend more time with the scholarship on Smith for the next draft of this chapter.]

Charlotte Smith is selected as a writer who was productive in multiple genres, only some of which may end up represented in corpora. Charlotte Smith's literary career began with the publication of her volume of poetry *Elegiac Sonnets*, in 1784. This work is the one upon which much of Smith's fame and prestige rested in the eighteenth century. A second edition of *Elegiac Sonnets* rapidly followed the first in the same year, with only slight amendments. The third and fourth editions of *Elegiac Sonnets* appeared in 1786, adding new poems. 1786 also saw the publication of Smith's *The Romance of Real Life*, a translation of *Les Causes Célèbres*, her first

foray into prose, which would occupy the major part of the next phase of her career. In 1788 she published her first original novel, *Emmeline, or the Orphan of the Castle*. 1789 begins this dissertation's decade of interest, a period of intense productivity for Smith: she had at least one new publication almost every year from 1789-1799. In 1789, she published her second original novel, *Ethelinde, or the Recluse of the Lake*, and a fifth edition of *Elegiac Sonnets*. In 1791 she published *Celestina*, her third novel; in 1792, her fourth novel, *Desmond*, and a sixth edition of *Elegiac Sonnets*. Although *Elegiac Sonnets* continued to be reprinted, reaching its tenth edition in 1812, after this edition no further poems were added. Instead, her new poetry appeared in their own independent publications, and no longer took the form of sonnets. In 1793 she published *The Emigrants*, a poem in two volumes, as well as *The Old Manor House*, her fifth novel. In 1794, her sixth and seventh novels, *The Wanderings of Warwick* and *The Banished Man*. In 1795 she published her eighth novel, *Montalbert*, and began writing in a new genre with *Rural Walks*. With *Rural Walks*, Smith's dominant genre again changed: having gone from a poet to a novelist, she now primarily published in a form which does not have a contemporary name: morally instructive natural history for "young persons." 1796 saw the sequel to *Rural Walks*, *Rambles Farther*, as well as the novel *Marchmont*, and the poem *A Narrative of the loss... of several ships*. 1797 saw the eighth edition of *Elegiac Sonnets*, unchanged since the sixth. 1798 saw the novel *The Young Philosopher*, and more natural history for children in *Minor Morals*. In 1799, Smith tried her hand at theatre with *What Is She?*, a comedy — not a form she will revisit. After this dissertation's decade of interest, Smith continued to write at a slightly less frenetic pace. In 1800 she published the first three volumes of *Letters of a Solitary Wanderer*, an epistolary anthology of narratives. In 1802 she published two additional volumes of *Letters of a Solitary Wanderer*. In 1804, she published *Conversations, Introducing Poetry*, for children. In 1806,

Lila Robinson!

Stalics

one!

Smith published *History of England*, another work for young persons, and Smith herself died, age 55. The next year saw the posthumous publication of the poem *Beachy Head* and the work for young persons, *The Natural History of Birds*.

Smith's personal life sometimes overshadows ^{her brilliant literary} ~~this~~ career. As her works often make clear to her readers, after a briefly comfortable youth as the daughter of a well-off country gentleman who lived beyond his means, she was married at age sixteen to Benjamin Smith, "son of a prosperous London merchant and owner of Barbados sugar cane plantations. The marriage was contracted hastily to remove her from her paternal home, now dominated by her new wealthy stepmother. Looking back in bitterness nearly forty years later, Charlotte Smith described the event as her father's decision to sell her like a 'legal prostitute, in my early youth, or what the law calls infancy' (Smith to Sarah Rose, 15 June 1804)" (Roberts). Benjamin Smith was cruel and ^{physically?} ~~violently~~ abusive. He was also so financially irresponsible that his wealthy father, Richard Smith, wanted to prevent Benjamin from inheriting. Charlotte Smith assisted Richard with business correspondence and impressed him as responsible and competent. In recognition of her husband's unreliability, "she persuaded [Richard] to relieve his son of all his ties to the business and establish him as a gentleman farmer in Hampshire" in 1774 (Zimmerman). Richard Smith died in 1776. "In an attempt to provide for his daughter-in-law, Richard bequeathed the bulk of his property to her children. But he had drawn up his will without professional advice; legal wranglings over the inheritance worth nearly £36,000 soon arose and were not settled until almost forty years later. By 1783 Benjamin had already unlawfully squandered more than a third of this trust and, as a consequence, found himself first in deep debt and then in King's Bench Prison." (Roberts). After the success of the *Elegiac Sonnets* allowed Smith to pay for her husband's release from prison, Benjamin Smith fled to France to escape further creditors.

Charlotte Smith moved between England and France over the next year and a half to negotiate his debts, and in 1785, the family was able to return to England. In 1787, after 22 years of marriage, Charlotte Smith legally separated from her husband, “an unusual step for a woman of her time” (Fry 7), and moved to a town near Chichester with her nine surviving children (of the twelve she had given birth to). However, despite this separation, Benjamin Smith retained a legal right to Charlotte Smith’s profits from her writing. Smith moved frequently after her separation, due to financial instability and declining health. “On 23 February 1806 Benjamin died in a debtors’ prison and some money reverted to Charlotte Smith. By then she was far too ill to execute her favourite scheme, to settle on the shores of Lake Lemane. On 28 October 1806 she died, only eight months after her husband, and seven years before Richard Smith’s estate was finally settled.” (Blank)

Smith’s posthumous critical reception has undergone multiple shifts in appreciation and obscurity. Duckling’s study of her presence in anthologies indicates that shortly after her death in 1806, Smith was widely eulogized and anthologized, remembered and emulated as an important British poet. As the nineteenth century went on, poetesses began to be anthologized separately from poets, in collections with ambitions that were ^{more} commercial ~~rather~~ than intellectual; Smith, too, “lost intellectual ground” even as she continued to be sold (Duckling 2016). By the end of the nineteenth century, even these volumes marginalized Smith’s poetry, with prefatory material which dismissed them as trite and depressing, unenjoyable reading. In the early twentieth century, Smith began to be ^{as} considered a novelist, rather than a poet; this new field did not lead at first to a much better reputation for her. Florence Hilbish produced the first extensive study of Smith, considering her as both poet and novelist, in 1941, to unappreciative reviews: Ernest Bernbaum’s faint praise said that “much time and care have been devoted to it; whether

?
Don't all
most coll.
have
commensurate
ends?

deservedly, is perhaps questionable,” since “the subtle or intricate is absent from Charlotte Smith’s writings” (138). Hilbish presents Smith’s emotional poetry as sincere rather than conventional, and her prose as more motivated by politics than commerce.

Duckling credits the feminist movement of the 1960s and 1970s with the beginning of Smith’s recovery (217): the renewed interest in women’s writing rediscovered her novels, and especially the radical political content which Hilbish had observed. At the same time, Bishop Hunt published a record of Smith’s influence on Wordsworth, as demonstrated by an almost overwhelming amount of physical evidence: Wordsworth owned copies of her works, which he annotated; he copied out some of her sonnets in his own hand; he paid her a personal visit; he edited some of her poetry for publication; he wrote explicitly of her influence in notes to his works. Hunt calls Smith “an important early influence on Wordsworth which has not been explored in any detail up to now” (85); his abstract somewhat snarkily asserts that “Wordsworth did not suddenly start writing sonnets in 1802 simply because he happened to read Milton’s.” However, Hunt has little praise for Smith herself: of one poem, he says, “Whatever the artistic value of such verses,” what matters is the underlying theme which Wordsworth would later express more masterfully (89). Smith continued to be treated separately as an interesting woman novelist, and a minor pre-Romantic poet, through the 1980s. Smith rose to greater prominence in both of these fields in the 1990s: with work by Stephen Curran, Roger Lonsdale, Jennifer Breen, Andrew Ashfield, and Jacqueline Labbe, “Smith became established not only as a prominent figure in the revised female canon, but also as a central figure in Romanticism” (Duckling 217).

Throughout this history, two aspects of Smith which have prompted frequent re-assessments are her personal life, and her work across genres. The first matter, the importance of a female author’s life as a woman to her importance as a figure worth remembering, is implicit in several

phases of the rise and fall described above. Fry is not alone in concluding that “[f]ew writers have presented themselves in their works so fully as did Charlotte Smith” (3): Smith’s poetry lyricizes her personal experiences, her novels feature autobiographical stand-in characters, and “the often intensely personal pleading prefaces” (Behrendt 189) to her works explicitly ask for them to be read light of her ongoing struggles. Perhaps as a result, much scholarship on Smith takes the stance of *The Literary Encyclopedia* in defining her as a woman who wrote because of, and chiefly about, her personal distress. Antje Blank’s article there highlights Smith’s financial motive to write: “Smith turned to writing when a failing marriage and a costly lawsuit left her without resources to raise her large family” (Blank). “And so,” Blank says, Smith “churned out” her novels (and the many editions of *Elegiac Sonnets*, and her other poetry, and her educational writing) to support herself and her nine children (Blank). Even when Smith’s *Elegiac Sonnets* “won her the reputation as an author of serious verse,” this is important primarily because it “lent greater respectability to her ensuing productions in a less prestigious but more lucrative genre – the novel” (Blank). At the same time, as Labbe argues in her article “Selling One’s Sorrows: Charlotte Smith, Mary Robinson, and the Marketing of Poetry,” Smith cultivated a public persona as a paragon of victimhood and motherhood, suffering deeply but turning her suffering into marketable prose out of a duty to her children. In periods where this image of womanhood is valuable, Smith is more easily valued, as in the eighteenth and nineteenth century anthologies which saw Smith as a moral exemplar (Duckling 203-4). Or, in periods when women’s resistance to patriarchal oppression is of scholarly interest, the direct, personal nature of Smith’s writing is valuable in itself, as in early feminist scholarship.



A complicating factor to these evaluations of Smith is that, as Labbe’s edited volume *Charlotte Smith in British Romanticism* thoroughly demonstrates, Smith’s writing is neither as

uniform nor as simplistically personal as autobiographical readings sometimes see it. Labbe contends that Smith-the-novelist and Smith-the-poet have been largely studied as separate entities, “and consequently we have been learning about two separate Smiths, each closely linked to the genre she writes in, neither closely linked to the other” (5). Labbe is not quite the first to attempt to unify Smith: Carol L. Fry’s 1996 monograph *Charlotte Smith* also addresses her poetry before moving on to the several phases of her novel-writing, including the children’s writing which made up much of Smith’s later career but does not appear in Labbe. Indeed, from the beginning, Hilbish’s 1941 monograph explicitly identifies Smith as “Poet and Novelist” in its title. However, Labbe is accurate regarding the somewhat different assessments of Smith current in the somewhat separate study of novels and of poetry in general: Labbe argues that as a novelist, Smith is now often praised for her innovative narrative techniques (implying a mode of writing that is intellectual and ‘distant’), whereas as a poet, she is praised for her innovative expressions of interiority (implying a mode of writing that is emotional and ‘close’). Labbe draws greater attention to important differences between Smith’s writing personae in different genres, and her edited collection “pulls together many Smiths” (2) to address these disjunctions. The volume not only addresses her novels and poetry, but also includes her plays, letters, and posthumous reception. Each of these Smiths, the volume contends, has something innovative and unexpected to reveal, important to the formation of British Romanticism. In Judith Phillips Stanton’s “Recovering Charlotte Smith’s Letters,” for example, Smith’s letters, less studied, reveal a third kind of writer, different from both the novelist and the poet, who conceives of herself as a professional businesswoman of her craft. More Smiths are available in genres not included in this volume, such as Smith the naturalist and children’s author (touched on only lightly in Labbe’s volume), or Smith the political philosopher who drives Amy Garnai’s

This
is
true

Revolutionary Imaginings in the 1790s, a highly political Smith who consciously participates in the “political public sphere” conceived by Habermas, despite Habermas’ insistence that women were excluded from this sphere (1). From these distinctions, Labbe concludes that “Smith, significantly, composes herself anew according to genre” (2) — and then asks, “Is it all to do with inherent qualities of genre, or is it more to do with the expectations we as readers bring to different genres?” (5). This question about genre is one of the initial questions to inspire this dissertation: to see it asked as a core question about Smith demonstrates Smith’s suitability as a figure whose career can shed light on important questions about the mediascape of the 1790s.

rearticulate in your own words

a little unclear. clarify?

1.4.2. Databases

A core object of study for this dissertation is the makeup and history of contemporary digital databases. Eighteenth-century materials of various kinds have been collected in many digital archives, of very different scopes. I will draw materials from the English Short-Title Catalogue (ESTC), Eighteenth Century Collections Online (ECCO), the ECCO Text Creation Partnership corpus (ECCO-TCP), Google Books, Project Gutenberg and HathiTrust. My examination of these six databases will, of necessity, examine a ‘time capsule’ of their holdings at a particular moment; the sources of my data, and my procedures for working with them, are described in more detail in Appendix B (“Methodology”). The databases vary from each other in terms of two main qualities: their size, and their reputation. The reputation of any given digital resource is shaped largely, I argue, by its ability to signal ‘rigour’ in its collection practices. Several databases of different sizes have established reputations of seriousness, and, correspondingly, cultural capital within scholarly communities. The databases that I will examine at length form two groupings of three each, to explore two sets of related concepts. The first set consists of ESTC, ECCO, and ECCO-TCP, all of which follow the same rigorous collection practices at

different scales. The second set consists of Google Books, HathiTrust, and Project Gutenberg, which follow very different collection practices while sharing a dubious scholarly reputation.

The first three databases I examine will be no surprise to eighteenth-century scholars: ESTC, ECCO, and ECCO-TCP. Gale's Eighteenth Century Collections Online (ECCO), contains over 180,000 titles 1701-1800, of which 42,000 were printed in England between 1789 and 1799. ECCO is itself (mostly) a subset of the broader English Short Title Catalogue (ESTC), which contains more 460,000 texts 1473-1800, of which 51,965 were printed in England between 1789 and 1799 (indicating that nearly 10,000 titles in the decade appear in the ESTC but not ECCO). The ESTC does not provide access to texts themselves: instead, it is an authoritative bibliographic catalogue, available as a searchable database. It is ECCO ^{which} provides texts: ECCO's 180,000 titles works are available as photographed facsimiles of the full text of each title. The facsimiles can be searched within ECCO's online interface; these searches examine a plaintext version of the facsimile pages that was generated by Optical Character Recognition (OCR), but this OCR text is not made directly available. As a result, the facsimiles may be read individually by scholars, but cannot form the basis for computational corpus analysis. A subset of ECCO's texts have been hand-prepared, as part of the Text Creation Partnership (TCP), to be easier to use in computational research. The resulting corpus of ECCO-TCP texts contains 2,231 titles, of which 466 were printed in England between 1789 and 1799. These titles are available as carefully edited texts encoded according to the Text Encoding Initiative (TEI) standard, which not only provides an accurate version of the text's words, but encodes substantial details regarding its context on the page. Most large scale distant reading of eighteenth-century literature relies on the ECCO-TCP corpus as its 'model' or 'sample' to represent the period. Accordingly, one of the tasks of this dissertation is to examine the makeup of this corpus, and how it differs

examples?

both from other corpora and from print culture in the period itself. These three digital collections — ECCO, ESTC, and ECCO-TCP — are the primary digital resources for the period, which form the basis of most digital research. However, they represent only one approach toward the collection and presentation of digital texts, to which there are two broad kinds of alternatives. These large but meticulous collections occupy a middle space between, on the one hand, highly selective thematic collections, such as The Shelley-Godwin Archive, of which there are many, and the giants of indiscriminate textual accumulation, such as Google Books, of which there are few.

Smaller collections allow for more scholarly curation, but have corresponding limitations. Whereas the ‘main players’ of the the mega-archives can be easily enumerated, these specialized collections are numerous. Some will focus on particular kinds of texts, such as the Early Novels Database (2,041 novels 1700-1799) or Broadside Ballads Online (more than 30,000 broadside ballads). Others exhaustively index particular publications, such as *The Hampshire Chronicle* (1,950 references to fiction in issues from 1772-1829), the Index to the *Lady's Magazine* (14,729 articles from 1770 to 1818), or the Novels Reviewed Database (1,836 reviews from *The Critical Review* and *The Monthly Review*, 1790-1820). Feminist scholarship in particular has seen the creation of resources like the Orlando Project, the Chawton House library Novels Online, Northeastern University's Women Writers Online and UC Davis's British Women Romantic Poets. The virtue of these collections is that they achieve even greater accuracy and comprehensiveness within their defined scope. The Shelley-Godwin Archive, for example, can reasonably aspire to digitize *every* known manuscript of Percy Bysshe Shelley, Mary Wollstonecraft Shelley, William Godwin, and Mary Wollstonecraft, and to provide these manuscripts in hand encoded plaintext transcripts. However, as is inevitable, these specialized

archives have the vices of their virtues: their specialized focus allows them to adapt precisely to their materials, and their idiosyncratic data structures can rarely be combined with other resources. The William Blake Archive, for example, benefits enormously from designing its archive around the unique images of each page of each copy of each of Blake's works. But because this approach is so well suited to Blake, it cannot be applied beyond Blake. Even if the archive's resources were available for download, they could not be directly compared to materials from another source which does not record its information at such a minute level of detail. As a result, although a great deal of excellent digital scholarship is contained in specialized micro archives, I do not examine them further in this dissertation.

Instead, I look at a set of larger archives of more contested "scholarly" status: Google Books, Project Gutenberg, and HathiTrust. Google Books may be the most infamous database of books. In a scholarly context, one hesitates even to designate this as an "archive," particularly in the same breath as resources like ECCO: books of all kinds are scanned indiscriminately with only the bare minimum of roughly accurate metadata collected about them. These rapidly scanned books are prone to unpredictable errors, including inaccurate dates, misspellings, duplicate copies, and inaccurate subject classifications¹⁹ — infamously, many books have "1899" assigned as their publication date because this date was used as a placeholder for "no date".²⁰ Nonetheless, Google Books is frequently used to study the prevalence of various "n-grams" (words or short phrases) over time, thanks to Google's built in tool. The tool is able to search books which are, for copyright restrictions, not available directly to readers, making it highly tempting for questions about contemporary language use.

Also in the category of smaller and specialized archives is Project Gutenberg. Project

¹⁹ (Harper 2016; Jacsó 2008; Weiss 2016) (CITE Mike Sutton and Mark D. Griffiths)

²⁰ CITE <http://languageolog.ldc.upenn.edu/nll/?p=1701>

in?

Gutenberg makes no claims to scholarly reliability but nonetheless underlies a not-significant amount of scholarly work — its cultural capital as a resource lags far behind its use and utility. Project Gutenberg is easily conceived of as a haphazard, ‘unscholarly’ source for materials, but unlike Google Books, Project Gutenberg actually does have selection criteria. Project Gutenberg will only collect public domain works which contemporary audiences might be interested in reading for pleasure. It narrows the field substantially to exclude works which have either ceased to be broadly interesting (as in the case of most forgotten fiction), or which were never particularly interesting (as in the case of almanacs and tax codes). Project Gutenberg includes 57,796 texts: far more than specialized scholarly archives like the Early Novels Database or the Shelley-Godwin Archive, but nonetheless an order of magnitude fewer than its more voracious potential competitors. And, like smaller specialized scholarly archives, Project Gutenberg has tailored its holdings to make it easy for readers to read, and quite difficult for its collection to be applied to any other use. By tailoring the structure of the archive itself to its specific materials, these collections are able to thoughtfully achieve their aims — but they also make it correspondingly difficult for users to achieve their own, different aims.

See p 9 36

is p. now!
COVID Access

What makes Google Books of interest in the context of this dissertation is its relationship to HathiTrust, an increasingly popular resource for scholars. HathiTrust’s collection contains digitized content from “a variety of sources, including Google, the Internet Archive, Microsoft, and in-house member institution initiatives.” The “in-house member institutions” include one hundred and fifty-five universities, colleges, and consortia of universities. The aggregate scholarly authority of these institutions carries the weight of elevating HathiTrust above the Google Books scans which form the backbone of much of its contents: “The members ensure the reliability and efficiency of the digital library,” the website assures us, “by relying on community

standards and best practices.” The texts themselves are stored in the database as facsimile page images and full-text OCR transcripts. In order to comply with copyright law, however, HathiTrust only provides large scale downloads and OCR transcripts for texts which are in the public domain. Most scholars use HathiTrust to run experiments on OCR transcripts of copyrighted texts, which they can only access through computational workarounds that intentionally make it impossible for the scholar to see the full transcript itself.²¹ These tools provide a unique solution to real barriers for digital scholars of contemporary literature: although copyright law would make it prohibitively expensive or even impossible to build corpora of post-1920s literature, HathiTrust’s mediated access to these texts enables corpus analysis. Through its collection, HathiTrust provides a hodgepodge of texts, of often unverifiable provenance and accuracy, selected largely by happenstance and convenience in a quest to contain all printed books. Through its tools, however, and through its institutional affiliations, HathiTrust has acquired a cultural capital among scholars which Google Books still lacks.

HathiTrust’s success in acquiring scholarly capital stands in interesting contrast with Project Gutenberg’s continued lack of cachet. Project Gutenberg is used in research with similar frequency to Google Books’ n-gram tool,²² but scholars often mention Project Gutenberg with a note of apology for not having found a better source. Its cultural capital as a resource lags far

²¹ For example, it might be able to acquire a text document with all of the words of a novel, but sorted into alphabetic order: such a text file can be used for some analyses based on word-frequency, but cannot be read. Or, it might be possible to find collocations of where a given word appears, but with only a limited number of words of context on either side of the term in question. Or, scholars can run pre-written code provided by HathiTrust to carry out things like topic modelling on the full, intact texts of their chosen works, but without being able to inspect those texts or run their own code on them. All of these modes of analysis make research much more difficult to carry out, and nearly impossible to verify. In the study of contemporary copyrighted literature, however, even these very limited tools for corpus analysis are valuable.

²² I have heard it quipped more than once in conference sessions that you always *think* that you’re going to get your texts from OCR, but you always *do* get them from Project Gutenberg.


behind its actual use and utility, likely, I argue, because its organizing principles are the 'unserious' ones of popularity and pleasure. Project Gutenberg is easily conceived of as a haphazard source for materials, but unlike Google Books, Project Gutenberg actually does have selection criteria. Project Gutenberg will only collect public domain works which contemporary audiences might be interested in reading for pleasure. This criteria might not render Project Gutenberg more useful for scholarly work but, it nonetheless narrows its selection substantially.

Project Gutenberg includes 57,796 texts: far more than specialized scholarly archives like the Early Novels Database or the Shelley-Godwin Archive, but an order of magnitude fewer than its more voracious potential competitors. In taking Project Gutenberg seriously as a collection of texts, I seek to explore the extent to which its reputation as "unreliable" may or may not be deserved.

As this brief survey of eighteenth-century digital archives shows, there is no 'perfect' corpus for large scale study of eighteenth-century texts. Moreover, I argue, the imperfect samples which each archive provides are shaped not only by historical factors of eighteenth-century print culture, but also by contemporary digital culture. Each archive represents a unique set of choices in response to the same sets of questions: what to include, why, how; what to make accessible, why, how, to whom; what, in the end, makes a text matter, and what we are meant to *do* with texts. As this dissertation will argue, these questions of digital history have important resonance with literary questions about literary canon formation.

1.5. Dissertation map

Chapter two describes in more detail the databases to be studied, and examines Charlotte Smith and the ways that her writing is made accessible today. The specific experimentation undertaken in chapter two tests the basic assumptions and methods of my project. I begin with a

the histories of the ESTC, ECCO, ECCO-TCP, Project Gutenberg, Google Books, and HathiTrust: highlighting the chronological relationships between these resources can explain each database's scope and technical implementation. Each new resource must contend with the possibility of either competing or collaborating with those which have come before. Examining materials like the meeting minutes and internal communications of these resources' early histories will show how those which currently enjoy the lowest reputation among scholars — Project Gutenberg and Google Books — defined their initial scope around an explicit rejection of scholarly norms. After establishing the history of each resource's development, I describe its current digital infrastructure, through the lens of critical algorithm studies. This begins with basic questions: what file formats does it use? What kinds of metadata, what ontologies? How does it make its materials available for use? Through close reading and comparison of these details, I articulate each database's implicit construction of what a text is and what it is for. Combining each resource's history with its technical infrastructure, I return to 

Having established these databases as objects of study, I identify what subset of Smith's works each corpus contains, as a concrete example to compare their holdings overall. Smith's *Elegiac Sonnets*, for example, are not included in the ECCO-TCP corpus (which is the one most often used for text mining research) — only *Celestina* and *The Emigrants* are included. Why these two texts? And what text mining research based on ECCO-TCP might have found slightly different answers if Smith's sonnets had been included? As a related test of comparison between databases, for each database which provides access to the actual text of Smith's works, I compare the textual similarity of *Celestina* and *The Emigrants*. What editorial choices are being made? How *much* worse is the OCR text than the transcribed text? Another key concept I will explore through Smith is the role of reprints. HathiTrust, for example, includes multiple editions of

Elegiac Sonnets. How reliable and effective are its distinctions between editions? How do the databases I examine handle multiple editions of a single work? I am particularly interested in how reprints can be incorporated into our understanding of what literature is “of” a particular decade: what does it mean to think of *Elegiac Sonnets*, initially printed in the 1780s, as “1790s literature”? Finally, having surveyed my six databases with the help of Smith, I discuss the multiple “Smiths” which emerge, and what it means to attempt to unify her disparate works.

In chapter three, I re-examine my core databases, but no longer with Smith as a focalizing lens. Instead, I undertake computational assessment and comparison of the databases’ contents. My research examines the authorship and subject matter (broadly construed) of all titles printed in England between 1789 and 1799 which are included in each database. I calculate the proportion of the titles in each resource that are attributed to men, to women, or are left unsigned. My naive hypothesis is that, as each resource demanded a greater investment of scholarly effort in each text, women and unsigned writers will grow increasingly underrepresented, so that the ECCO-TCP corpus will have substantially different demographics than the ESTC. Using the titles of these works and a topic modelling tool which I have built, I also roughly identify the subject matter of each title, categorizing works into broad genres such as drama, poetry, Romance, History, or sermons. Although the topic modelling tool is able to cluster what it sees as “similar” titles, individual interpretation is required to make these clusters meaningful. A substantial portion of chapter three is dedicated to discussing how scholars apply genre categories retrospectively to clusters of texts, how publishers sought to advertise their texts to particular audiences, and how the categories I develop ought to be understood in the context of existing eighteenth century scholarship on print genres.²³ Using my resulting genre

²³ I am particularly excited to explore “false advertising” in titles!

classifications, I am then able to compare these four resources to each other, and to existing scholarly work on the print production of the 1790s. For example, I compare each resource's holdings to the statistics on the English novel included in Garside, Raven and Schöwerling's *Bibliographical Survey of Prose Fiction*. In examining genres, I anticipate discovering a preference, in the more specialized resources, for more "literary" forms of writing.

I will also correlate gender and genre. This preference for particular kinds of writing might explain changes in gender representation of smaller corpora. If the novel is the domain of women, for example, a corpus can underrepresent women by underrepresenting novels. Or it could include a representative number of novels, but disproportionately include novels by men. My investigation allows me to identify the patterns of exclusion. Asking bibliographical questions of multiple corpora, in order to learn about the corpora themselves, emphasizes an under-examined stage of text mining research, and provides a basis for other scholars to use these corpora more precisely. It also provides the foundation for my own

In my fourth chapter, I playfully attempt what might be considered a devil's advocate method of textual selection: pure random sampling. Using a random number generator, I select arbitrary texts to close read, and weave together a narrative of 1790s print from their contents. Much of my work will involve defining and justifying the parameters for my random selection — ESTC, or a full-text database? How many texts? From which years? — but once I have taken my sample, I will not re-sample. For each text, I explore the path which brought it into the databases in question, and what scholarship (if any) might be used to interpret it. How far afield do I have to look, to find scholarly conversations addressing each text? What, if anything, can be produced by placing them in conversation? This methodology is inspired by work in the field of speculative computing, a practice of creating strange and possibly non-functional programs in order to generate productive forms of surprise.

A final brief conclusion to this dissertation offers an assessment of the role of digital textual collections in contemporary literary study.

Chapter Two

2.1. Databases

In the next section I will close-read the implicit models underlying each database, to examine how each enforces a particular concept of “literature” and “a text.” However, before these models can make sense, we must understand the history of how they were built. I contend that each database is best understood as a negotiation between commercial and noncommercial values. Each database has the goal of making valuable information available. After the 1990s, they are particularly influenced by the utopian ideal that digital reproduction at last made textual reproduction free. Each had to contend, however, with the fact that before a text can be reproduced digitally it must be created digitally, and that even if the material costs are entirely eliminated (which, of course, they are not) textual creation continues to have costs in labour.

[The histories of these databases currently run to more than 9,000 words of bare recitations of facts. My next steps are to further clarify some of the details, and to frame and narrativize the information to make it meaningful. A brief timeline of milestones follows.]

- 1918 Pollard first proposes a “short-title handlist”
- 1926 Pollard and Redgrave Short-Title Catalogue for 1476–1640
- 1938 Eugene B. Power founds University Microfilms
- 1945 Wing starts collecting his STC, 1641–1700
- 1951 Donald Wing’s catalogue for 1641–1700, first edition
- 1971 First text in what would be Project Gutenberg. Over the next twenty years, Michael Hart personally keyed the first hundred books.
- 1972 Beginning of second ed of Wing STC, 1641–1700
- 1976 Proposal for Eighteenth Century Short Title Catalogue, British Library and the American Society for Eighteenth Century Studies

- 1976 Second edition, vol 1, of Wing's STC
- 1976 Beginning of second ed of Pollard & Redgrave STC, 1475-1640
- 1977 ESTC pilot begun at British Library, directed by Robin Alston
- 1979 ESTC: Libraries from USA, Germany, and Australia began contributing to ESTC
- 1980 ESTC database available via British Library BLAISE [British Library Automated Information Service]
- 1981 Research Publications, Inc begins microfilming books
- 1981 ESTC database available via US Research Libraries Group RLIN [Research Libraries Information Network] system
- 1983 ESTC catalogue of BL holdings and indexes published in microform
- 1983 *Eighteenth Century Collection* microfilm produced by Research Publications, Inc
- 1985 ESTC online databases in RLIN and BLAISE upgraded to allow dynamic updates to a single shared file
- 1986 Second edition, vol 2, of Wing's STC
- 1987 ESTC expanded scope to add all print prior to 1700, changing its name to the English Short Title Catalogue. Information from Wing and STC is added to ESTC.
- 1987 Michael Hart recruits first Project Gutenberg volunteers
- 1989 Project Gutenberg completes its tenth book, the King James Bible
- 1991 End of second edition of Pollard & Redgrave STC, 1475-1640
- 1991? Exhaustive index to Wing's STC — after which Bibliographical Society no longer supported Wing
- 1992 ESTC expanded scope to add serials
- 1994 ESTC made pre-1700 records available
- 1994 Project Gutenberg completes its 100th book, the Complete Works of William Shakespeare ?? (Contradicts 1987 count)
- 1994 Project Gutenberg's first website is developed by volunteer Pietro Di Miceli

“By the late 1990s, several thousand reels had been published in two series: ‘Early English Books, 1475–1640’ and ‘Early English Books, 1641–1700’.” (Gadd)

1997 Project Gutenberg publishes its 1000th book, *La Divina Commedia di Dante*, in Italian

1998 ESTC second edition released on CD-ROM

1998 Conclusion of second ed of Wing STC

1998 Beginnings of EEBO: University Microfilms (now ProQuest) began to make available digitised copies of its microfilms across the Internet to subscribing institutions

1999 ESTC assumed official responsibility for receiving new Wing STC data

1999 TCP began

2000 Project Gutenberg: Charles Franks launches Distributed Proofreaders

2003 ESTC third edition released on CD-ROM

2003 Project Gutenberg 600 “best” ebooks released on CD-ROM, followed by 10,000 item DVD

2003 Beginning of ECCO: Thomson Gale (now Gale Cengage Learning) made digital copies of Eighteenth Century Collection microfilms available to subscribers online

2004 Google Print is announced

2005 TCP begins encoding ECCO texts

2006 ESTC made available to search free online; ESTC begins transcribing full title and imprints

2007 Project Gutenberg DVD released with 17,000 items

2008 Project Gutenberg publishes its 25,000th book, *English Book Collectors*, by William Younger Fletcher

2008 HathiTrust founded, by 12-university Committee on Institutional Cooperation and 11-library University of California Libraries

2009 EEBO-TCP Phase I complete: produced 25,000 books; beginning of Phase II

2010 Project Gutenberg DVD released with 30,000 items

- 2011 40,000 books in Project Gutenberg
- 2015 EEBO-TCP Phase I books released to the general public
- 2017 Project Gutenberg discontinues free mailing of CDs and DVDs, though the files remain available for people to burn their own copies at home
- 2021 EEBO-TCP Phase II books released to the general public

2.2. Charlotte Smith

For the purposes of this chapter, I examine Smith's works which fall outside this dissertation's decade of interest. As Table 1 shows, Smith's publishing career began in 1784 and continued until her death in 1806; when I refer to Smith's "full" output, I consider all 47 editions of her works published in her lifetime or in the year immediately following her death. Her 1790s output (that is, the editions published 1789-99) consists of 30 of those editions. I have slightly expanded my chronological focus in part because some of the most interesting exclusions occur earlier and later in Smith's publishing career, such as the first edition of her immensely influential *Elegiac Sonnets* (1784), which is listed in the ESTC but not available in facsimile anywhere, or the publications in the last years of her life, which are excluded from the chronological focus of most resources but can still appear in HathiTrust. Of particular interest is the fact that *Beachy Head*, which is now one of Smith's most frequently anthologized and taught poems, does not appear in a single digital database. None of these inclusions or exclusions represent an agenda against (or for) Smith, or indeed an interpretive choice at all, but they nonetheless shape the disciplinary infrastructure.

year	title	ed	ESTC	ECCO	ECCO-TCP	Hathi
1784	Elegiac Sonnets, vol 1	1st ed	ESTC yes	ECCO no	TCP no	Hathi no
1784	Elegiac Sonnets, vol 1	2nd ed	ESTC yes	ECCO yes	TCP no	Hathi no
1788	Elegiac Sonnets, vol 1	3rd ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1787	Romance of Real Life	1st ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1788	Emmeline	1st ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1788	Emmeline	2nd ed	ESTC yes	ECCO yes	TCP no	Hathi no
1789	Ethelinde	1st ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1789	Emmeline	3rd ed	ESTC yes	ECCO yes	TCP no	Hathi no
1789	Elegiac Sonnets, vol 1	5th ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1790	Ethelinde	2nd ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1791	Celestina	1st ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1791	Celestina	2nd ed	ESTC yes	ECCO yes	TCP yes	Hathi yes
1792	Desmond	1st ed	ESTC yes	ECCO yes	TCP no	Hathi no
1792	Desmond	2nd ed	ESTC yes	ECCO yes	TCP no	Hathi no
1792	Elegiac Sonnets, vol 1	6th ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1793	The Old Manor House	1st ed	ESTC yes	ECCO yes	TCP no	Hathi no
1793	The Emigrants	1st ed	ESTC yes	ECCO yes	TCP yes	Hathi yes
1793	The Old Manor House	2nd ed	ESTC yes	ECCO yes	TCP no	Hathi no
1794	The Banished Man	1st ed	ESTC yes	ECCO yes	TCP no	Hathi no
1794	Wanderings of Warwick	1st ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1795	Rural Walks	1st ed	ESTC yes	ECCO yes	TCP no	Hathi no
1795	Montalbert	1st ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1795	Rural Walks	2nd ed	ESTC yes	ECCO yes	TCP no	Hathi no
1795	The Banished Man	2nd ed	ESTC yes	ECCO no	TCP no	Hathi yes
1795	Elegiac Sonnets, vol 1	7th ed	ESTC yes	ECCO yes	TCP no	Hathi no
1796	A Narrative of the loss...	1st ed	ESTC yes	ECCO yes	TCP no	Hathi no
1796	Rambles Farther	1st ed	ESTC yes	ECCO yes	TCP no	Hathi no
1796	Marchmont	1st ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1797	Elegiac Sonnets, vol 2	1st ed	ESTC yes	ECCO yes	TCP no	Hathi no
1797	Elegiac Sonnets, vol 1	8th ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1798	Minor Morals	1st ed	ESTC yes	ECCO no	TCP no	Hathi no
1798	The Young Philosopher	1st ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1798	Rural Walks	3rd ed	ESTC yes	ECCO yes	TCP no	Hathi no
1799	What Is She?	1st ed	ESTC yes	ECCO yes	TCP no	Hathi no
1799	Minor Morals	2nd ed	ESTC yes	ECCO yes	TCP no	Hathi no
1799	What Is She?	2nd ed	ESTC yes	ECCO yes	TCP no	Hathi no
1800	Letters of a Solitary Wanderer, vols 1-3	1st ed	ESTC yes	ECCO yes	TCP no	Hathi yes
1800	Elegiac Sonnets, vol 1	9th ed	ESTC yes	ECCO no	TCP no	Hathi no
1800	Elegiac Sonnets, vol 2	2nd ed	ESTC yes	ECCO yes	TCP no	Hathi no
1800	What Is She?	3rd ed	ESTC yes	ECCO yes	TCP no	Hathi no
1800	Rural Walks	4th ed	ESTC yes	ECCO yes	TCP no	Hathi no
1800	Rambles Farther	2nd ed	ESTC yes	ECCO no	TCP no	Hathi no
1802	Letters of a Solitary Wanderer, vols 4-5	1st ed	ESTC no	ECCO no	TCP no	Hathi yes
1804	Conversations, Introducing Poetry	1st ed	ESTC no	ECCO no	TCP no	Hathi no
1806	History of England		ESTC no	ECCO no	TCP no	Hathi no
1807	Beachy Head	1st ed	ESTC no	ECCO no	TCP no	Hathi no
1807	Natural History of Birds	1st ed	ESTC no	ECCO no	TCP no	Hathi no

Table 1: All editions of Charlotte Smith's works published in England during her lifetime or in the year immediately following her death, and their inclusion in the ESTC, ECCO, ECCO-TCP, and HathiTrust databases.

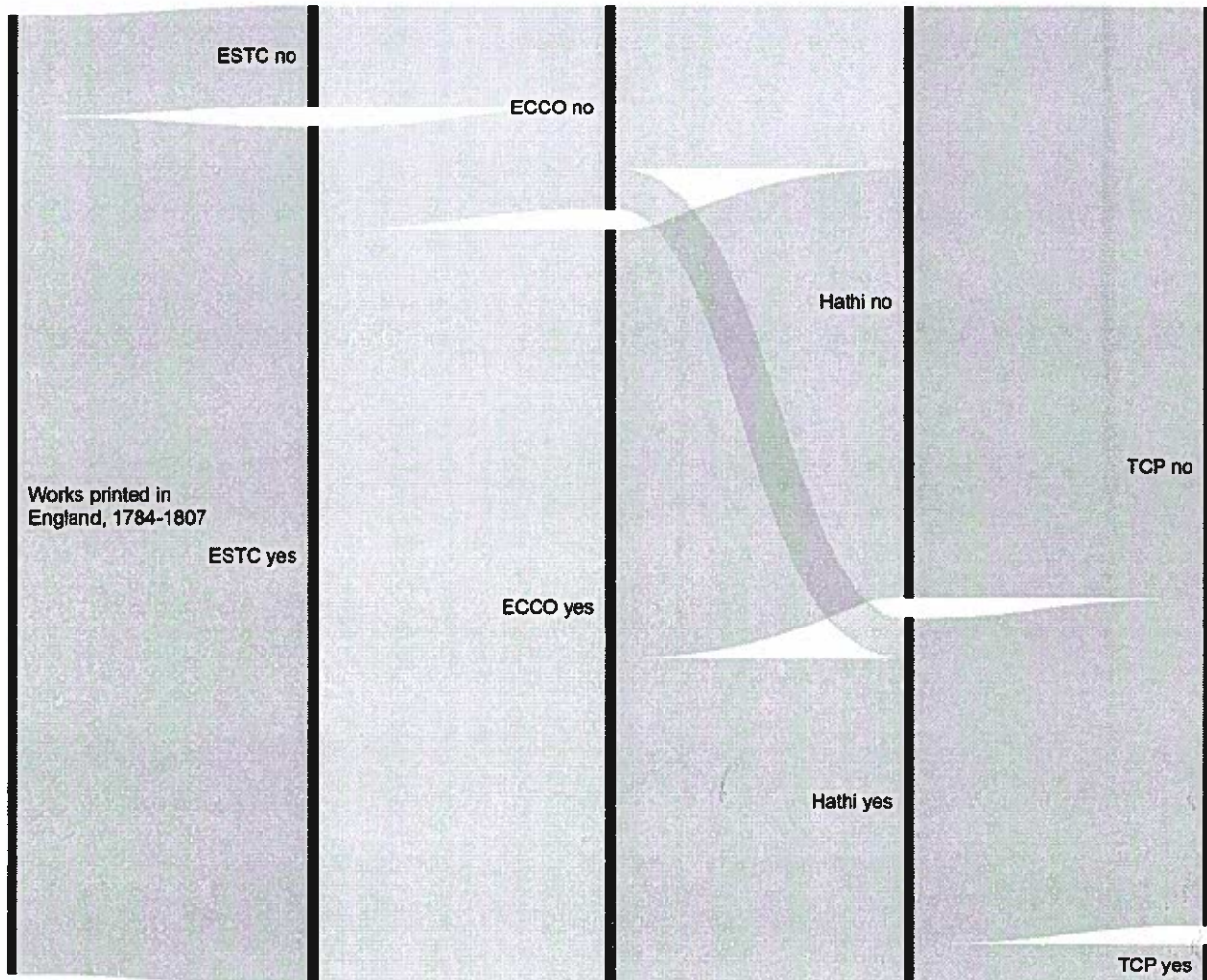


Figure 1: An alluvial chart, showing the winnowing down of Smith's works from database to database. Of the 47 editions printed in England between 1784 and 1807, 42 are included in the ESTC, and 5 do not appear in the ESTC because they were printed after 1800 and thus fall outside its purview. ECCO contains 37 of Smith's 47 editions, all of which also appear in the ESTC. ECCO is missing the 5 editions not listed in the ESTC (since it, too, does not contain works past 1800), as well as another 5 works. HathiTrust contains 18 of Smith's 47 editions, but unlike ECCO, these are not a simple subset of the ESTC. HathiTrust contains one of the 5 editions excluded from the ESTC, and one of the 5 editions included in ESTC but excluded from ECCO. The remaining 16 HathiTrust editions appear in both the ESTC and ECCO. ECCO-TCP includes only 2 of Smith's 47 editions, both of which appear in every previous database. Graph generated using RAW Graphs (Mauri et al.).

Figure 1 shows how Smith's presence in four major databases has the effect of winnowing

down her full output arbitrarily. Even the largest collection, the 42 editions included in the ESTC, is not comprehensive: since the ESTC does not include any works published after 1800, it excludes volumes 4 and 5 of *Letters of a Solitary Wanderer* (1802), three works for children (*Conversations*, *Introducing Poetry*, 1804; *History of England*, 1806; and *Natural History of Birds*, 1807), and the posthumous publication that now forms a major part of Smith's reputation as a poet, *Beachy Head* (1807). ECCO lacks these five editions for the same reason, and is also missing five others: the first and ninth editions of *Elegiac Sonnets* (1784 and 1800), the second edition of *The Banished Man* (1795), the first edition of *Minor Morals* (1798), and the second edition of *Rambles Farther* (1800).

HathiTrust contains 18 of Smith's 47 editions, though these are not a simple subset of the ESTC and ECCO. Unlike the ESTC and ECCO, HathiTrust contains volumes 4 and 5 of *Letters of a Solitary Wanderer* (1802)²⁴. This is the only post-1800 work which appears in HathiTrust, however—the others are also missing, including the important volume *Beachy Head* (1807). There is one work included in HathiTrust but not in ECCO, the second edition of *The Banished Man* (1795). Whereas ECCO does not include works unless there is a complete copy available, HathiTrust provides scans of volumes 2, 3, and 4, and simply implies through their numbering that there is a missing first volume — perhaps in the optimism that a volume 1 will appear from another library's holdings, to complete the set later.²⁵ The remaining HathiTrust included titles appear in both the ESTC and ECCO, and a further 21 titles appear as facsimiles in ECCO but not

²⁴ Volumes 4 and 5 of *Letters of a Solitary Wanderer* are in fact part of the same bibliographic record as the first three volumes. The publication date for the combined five-volume work is listed as "1800-1802."

²⁵ Several of HathiTrust's records provide "mixed copies" like this, with some volumes scanned from one library's holdings and other volumes scanned at another. If there is overlap, multiple scans will be provided for the duplicated holdings. Nonetheless, all of these scans are tied to a single unified MARC record, taken from only one of the holding library (with no indication of which library provided it).

in HathiTrust. At first blush it is somewhat surprising that HathiTrust has failed to include works which are, demonstrably, in known locations at institutional libraries, and in physically sound condition to be scanned—but the scans making up HathiTrust bear no relation to the scans in ECCO. *The Young Philosopher* (1798), for example, appears in ECCO sourced from a British Library copy, but the HathiTrust images are “Google-digitized” from the New York Public Library. Google’s rapacious book-scanning, evidently, was not as thorough as ECCO’s sustained scholarly project.

The smallest subset of all of these texts is the ECCO-TCP holding of just two titles: the second edition of *Celestina* (1791), and the first edition of *The Emigrants* (1793). Both titles appear in all larger databases, including HathiTrust (though, as I will discuss, they arrive in HathiTrust from a different source). *The Emigrants* is included in ECCO-TCP as one file, based on the ECCO facsimile of an original from the Huntington Library. *Celestina* is included as four files, one for each of four volumes, based on the ECCO facsimile of an original from the British Library. Both works were first reproduced in the microfilm version produced 1982-2002 in by Research Publications,²⁶ then digitized in 2003 (released on ECCO in June 2004), and finally published as TEI XML files in January 2007. The current files have been kept up to date with changes in TEI standards, and were created by converting TCP files to TEI P5 using tcp2tei.xsl. The bibliographic metadata for these works is the same between ESTC, ECCO, and ECCO-TCP records. In HathiTrust, however, the source text for *The Emigrants* is a University of California Library copy, rather than the British Library, scanned by Google Books, and presented with substantially less detailed bibliographic information. The ESTC, ECCO, and ECCO-TCP records for *The Emigrants* all provide the same physical description “ix,[3],68[i.e. 60]p. ; 4⁰” with the

²⁶ Later known as Primary Source Microfilm, an imprint of the Gale Group.

same note “[n]umbers 9-16 omitted in pagination; text is continuous.” HathiTrust, in contrast, gives the physical description “ix, 68 p. ; 26 cm,” which is both more and less information: a quarto volume could be a range of sizes, so HathiTrust provides new detail by giving a measurement in centimetres, but the data on page numbers is now misleading. Consulting the HathiTrust facsimile shows that it, too, omits the page numbers 9-16, going directly from page 8 to page 17 without a break in the poem. HathiTrust also omits information on the three unnumbered pages between the preface and the poem. Evidently, a human did consult the book, to identify a nine-page preface in roman numerals, and the page number on the last page, but they did not carry out a full collation.

Only one of Charlotte Smith’s works is available in Project Gutenberg: *Emmeline, the Orphan of the Castle* (first published 1788).

Searching the ESTC for records which both have “Toronto” in the library name and “Charlotte Turner” in the author name turns up two records: volume one of *Rural Walks* (1795) and *Minor Morals* (1798), both held at the Toronto public library. The ~~Toronto Public Library~~ ^{TPL} ~~OK~~ catalogue has two distinct author identities for “Smith, Charlotte Turner, 1749-1806, author.” and for “Smith, Charlotte, 1749-1806,” and the special collections holdings only appear under the latter name (making them initially difficult to find). Under the “Smith, Charlotte” name, however, six titles printed during Smith’s appear: the two listed in ESTC, plus a complete two-volume copy of *Rural Walks* (1795), the first and second editions of *Rambles Farther* (1796 and 1800), and *Conversations Introducing Poetry* (1804). Of these, *Rural Walks* and both editions of *Rambles Farther* are listed in the ESTC but without records of the Toronto copies. All six titles are part of the Osborne Collection of Early Children’s Books. This is interesting because it shows how scholarly disciplinary interpretations perpetuate themselves *infrastructurally*: as a Toronto-

✓
TPL's

based scholar, the path is easier for me to study Smith-the-children's-writer than other Smiths.

[The final section on Smith is a survey of some distant-reading projects which made use of these various databases, with discussion of which works by Smith would have been included (or not) in their corpora.]

2.3. OCR

How accurate does OCR need to be? This depends on how the OCR will then be used. There is a substantial body of existing literature on this subject, which I will survey and integrate here. Thinking in these terms offers a different and sometimes uncomfortable definition for what it might mean to call a text “reliable”. In addition to evaluating the reliability of available OCR copies of Smith’s works, I will evaluate Project Gutenberg’s reliability. Project Gutenberg makes no guarantees of matching any particular edition of a text, and often presents Victorian editions (potentially bowdlerized) without warning, but nonetheless provides texts which are entirely made of actual words: does this make them more, or less, “reliable” than OCR? How can different research questions change that evaluation?

The existence of a carefully hand-corrected transcription of *The Emigrants* in ECCO-TCP provides an opportunity to check the reliability of the OCR in both ECCO and HathiTrust. I will proceed from the assumption that the ECCO-TCP files are 100% accurate, and that any differences between the OCR and ECCO-TCP represents an OCR error.²⁷ Before beginning the experiment, my hypothesis was that both ECCO and HathiTrust would differ from each other in where and how they are inaccurate, but would have similar accuracy overall. I suspected that

²⁷ One exception to this assumption has to do with treatment of the character f, which the TCP file modernizes to an s, but which HathiTrust renders as f. To avoid penalizing HathiTrust for “inaccuracy” when it is actually a more accurate reproduction of the page than my reference point, I amended every instance of f in HathiTrust to an s.

they were likely around 50% accurate, plus or minus 10% — I wouldn't be surprised if they were worse, but would be quite surprised if their accuracy was 80% or higher. Acquiring the plaintext files from all three sources required some hunting for some hidden options and some workarounds; rendering them suitable for comparison required some modifications of each file, described more fully in Appendix B. Although Gale Digital Scholar Labs prominently provided an “OCR Confidence” of 95%, the first glance at the document was not very promising. To my surprise, Juxa calculated a relatively low “change index” for each text compared to the TCP witness: ECCO had a .16 change from base (i.e., 84% accuracy), and my normalized HathiTrust document had only a .09 change from base (i.e., 91% accuracy).²⁸ This surprised me, and suggests that skepticism of OCR in eighteenth century text mining may no longer be appropriate.

To make these comparisons concrete, consider the first page of Smith's dedication, as it is captured by OCR in ECCO and HathiTrust, and in the ECCO-TCP transcript:

TO WILLIAM COWPER, Es DEAR SIR, THERE is,- I hope, some propriety in my
addrefing a Com- potion to you, which would,never perhaps have existed, had I not, amid
the heavy prefure of many sorrows, derived infinite consolation from your Poetry, and
some degree of animation and of confidencefrom your efieen. . 'he.following
performance isfarfrom aspiring to be con- fidered as an imitation of your inimitable
Poem, " THE " TASK;" I am perfectly sensible, that it belongs not to a feeble
andfemninine hand to draw the Bow of Ulyfes.,Theforce, clearness, and sublimity ofyour
admirable Poem; the felicity, almost peculiar to your genius, of givingto the moJ familiar
objegls dignity and eset, I could never hope to,a reach (ECCO)

T O WILLIAM com/PER, Ess. DEAR SIR, THERE is, I hope, some propriety in my
addreffing a Com- position to you, which would never perhaps have existed, had I not,
amid the beavy preffure of many forrows, derived infinite consolation from your Poetry,
and some degree of animation and of confidence from your °fteem. The following
performance is far from aspiring to be con- fidered as an imitation of your inimitable
Poem, “ The “TAsk;” I am perfy f°ol, that it belongs not to a feeble and feminine band to

²⁸ Leaving the f characters unchanged in the HathiTrust document resulted in a .29 change from base (71% accuracy), so my normalization of f to s had a major impact on the comparison. I consider the .09 result more appropriate than the .29 because the normalized copy better reflects how an OCR file would be used.

draw the Bow of Ulysses. The force, clearness, and sublimity of your admirable Poem; the felicity, almost peculiar to your genius, of giving to the most familiar objects dignity and effect, I could never hope to 3. - Reach (HathiTrust)

TO WILLIAM COWPER, ESQ.

DEAR SIR,

THERE is, I hope, some propriety in my addressing a Composition to you, which would never perhaps have existed, had I not, amid the heavy pressure of many sorrows, derived infinite consolation from your Poetry, and some degree of animation and of confidence from your esteem.

The following performance is far from aspiring to be considered as an imitation of your inimitable Poem, "THE TASK;" I am perfectly sensible, that it belongs not to a feeble and feminine hand to draw the Bow of Ulysses.

The force, clearness, and sublimity of your admirable Poem; the felicity, almost peculiar to your genius, of giving to the most familiar objects dignity and effect, I could never hope to (ECCO-TCP)

Both of the OCR copies contain errors in individual letters which render the whole word interpretable by a human but not by text mining software, as in the case of "beavy" for "heavy." The ECCO copy struggles with the fact that l is not an available character, sometimes substituting an f, as in "prefure" for "pressure." Both leave out spaces between words, creating new tokens like "isfarfrom" and "andsublimity," though HathiTrust is less prone to this error.

Other features of the OCR copies are accurate to the page image but would nonetheless interfere with text mining. The hyphenation of "Com- position," for example, would prevent it from rendering as a single word, though here even the careful TCP copy would introduce the same problem, since the line break is encoded as "Com|position." Before the TCP copy could be used for text mining, the | characters would likely need to be removed — not too different from

removing the hyphenation from the ECCO and Hathi copies. Most difficult to resolve is the fact that OCR naturally attempts to capture *all* text on the page, including the signature mark and catch word. In ECCO these appear as “,a reach” and in Hathi they are “3. - Reach” whereas TCP more appropriately leaves these out. Unlike the problems with hyphenated words, there is no way to correct for the inclusion of catchwords in a document, since there is no predictable way to identify them — but keeping them in the document will cause any text-mining software to count these words twice.

The usual “text cleaning” procedures would further prepare these OCR texts for text mining by transforming all words to lowercase, removing all punctuation, and, in most cases, deleting all words which don’t match a predefined dictionary of valid words. A scholar working with the HathiTrust OCR would almost certainly add to this a step converting the f character to an s, as discussed above, in order to make the dictionary comparison feasible. The result of this ‘cleaning’ would likely look something like the following:

to william dear sir there is i hope some propriety in my a potion to you which would never perhaps have existed had i not amid the heavy of many sorrows derived infinite consolation from your poetry and some degree of animation and of your he following performance aspiring to be con as an imitation of your inimitable poem the task i am sensible that it belongs not to a feeble hand to draw the bow of clearness and sublimity admirable poem the felicity almost peculiar to your genius of the familiar dignity and i could never hope to a reach (ECCO, as it would likely appear after text “cleaning”)

william dear sir there is i hope some propriety in my addressing a position to you which would never perhaps have existed had i not amid the pressure of many sorrows derived infinite consolation from your poetry and some degree of animation and of confidence from your the following performance is far from aspiring to be considered as an imitation of your inimitable poem the task i am that it belongs not to a feeble and feminine band to draw the bow of ulysses. the force of your admirable poem the felicity almost peculiar to your genius of giving to the most familiar dignity and i could never hope to 3 reach (HathiTrust, as it would likely appear after text “cleaning”)

Strikingly, these ‘clean’ texts are now further from legible to human eyes, as OCR errors which a reader could mentally correct (such as “beavy” for “heavy” are now entirely removed.

TO WILLIAM COWPER, ESQ.
DEAR SIR,

THERE is, I hope, some propriety in my addressing a Composition to you, which would never perhaps have existed, had I not, amid the heavy pressure of many sorrows, derived infinite consolation from your Poetry, and some degree of animation and of confidence from your esteem.

The following performance is far from aspiring to be considered as an imitation of your inimitable Poem, "THE TASK;" I am perfectly sensible, that it belongs not to a feeble and feminine hand to draw the Bow of Ulysses.

The force, clearness, and sublimity of your admirable Poem; the felicity, almost peculiar to your genius, of giving to the most familiar objects dignity and effect, I could never hope to

Figure 2: Juxta’s “Heat Map” visualization of the “base” witness of the first page of *The Emigrants* (i.e., the ECCO-TCP version carefully prepared by scholars), highlighting words which differ in the two witnesses of the ECCO OCR and the normalized HathiTrust OCR. A darker highlight indicates that the word varies in more than one witness.



Figure 3: A histogram, produced by Juxta, showing where the two ECCO and normalized HathiTrust witnesses show the most difference from the base ECCO-TCP copy. “Longer lines indicate areas of considerable difference, while shorter lines indicate greater similarity between documents.” (“A User Guide to Juxta Commons”)

Works Consulted

- Algee-Hewitt, Mark. “Acts of Aesthetics: Publishing as Recursive Agency in the Long Eighteenth Century.” *Romanticism and Victorianism on the Net*, vol. 57-8, 2010, doi:10.7202/1006517ar.
- Alston, Robin. “The Eighteenth Century Short Title Catalogue: A Personal History to 1989.” web.archive.org/web/20080908103158/http://www.r-alston.co.uk/estc.htm.
- Bainbridge, Simon. *British Poetry and the Revolutionary and Napoleonic Wars: Visions of Conflict*. Oxford UP, 2003.
- Baldick, Chris, and Robert Mighall. “Gothic Criticism.” *A New Companion to The Gothic*, edited by David Punter, Wiley-Blackwell, 2012, pp. 265-287, doi:10.1002/9781444354959.ch19.
- Barthes, Roland. “The Reality Effect.” 1968. *The Rustle of Language*, translated by Richard Howard, Hill and Wang, 1986, pp. 141-148.
- Baskin, Jon. “On the Hatred of Literature.” *The Point*, issue 21, 26 January 2020, [/web/20200506015431/https://thepointmag.com/letter/on-the-hatred-of-literature/](https://thepointmag.com/letter/on-the-hatred-of-literature/).
- Battis, Jes. “Molly Canons: The Role of Slang and Text in the Formation of Queer Eighteenth-Century Culture.” *Lumen*, volume 36, 2017, pp. 129-141. doi:10.7202/1037858ar.
- Bauder, Julia. “HathiTrust as a Data Source for Researching Early Nineteenth-Century Library Collections: Identification, Coverage, and Methods,” *Information Technology and Libraries*, volume 38, issue 4, December 2019. ProQuest, ProQuest document ID 2336298791.
- Bayard, Pierre. *How to Talk About Books You Haven't Read*, translated by Jeffrey Mehlman. Bloomsbury, 2007.
- Behrendt, Stephen C. “Charlotte Smith, Women Poets and the Culture of Celebrity.” *Charlotte Smith in British Romanticism*, edited by Jacqueline Labbe. Pickering & Chatto, 2008, pp. 189-202.
- Benatti, Francesca and Justin Tonra. “English Bards and Unknown Reviewers: A Stylometric Analysis of Thomas Moore and the *Christabel* Review.” *Brea: A Digital Journal of Irish Studies*, 7 Oct 2015,

- web.archive.org/web/20191107181606/https://breac.nd.edu/articles/english-bards-and-unknown-reviewers-a-stylometric-analysis-of-thomas-moore-and-the-christabel-review.
- Bernbaum, Ernest. Review of Charlotte Smith, Poet and Novelist (1749-1806) by Florence May Anna Hilbish. *Modern Language Notes*, vol. 59, no. 2, 1944, pp. 137–139. JSTOR, www.jstor.org/stable/2910610.
- Blanch, Anna Maree. A Reassessment of the Authorship of the Cheap Repository Tracts. Master's thesis, Baylor University, 2009.
- Blank, Antje. "Charlotte Smith." Edited by Janet Todd. *The Literary Encyclopedia*, volume 1.2.1.06: English Writing and Culture of the Romantic Period, 1789-1837, edited by Daniel Cook and Daniel Robinson, 23 June 2003, www.litencyc.com. Accessed 05 June 2019.
- Blaney, Jonathan. "Introduction to the Principles of Linked Open Data." *The Programming Historian*, volume 6, 2017, web.archive.org/web/20200228212040/https://programminghistorian.org/en/lessons/intro-to-linked-data. Accessed 28 February 2020.
- Blayney, Peter W. M. "The Alleged Popularity of Playbooks." *Shakespeare Quarterly*, vol. 56, no. 1, 2005, pp. 33–50. JSTOR, www.jstor.org/stable/3844025.
- Blevins, Cameron and Lincoln Mullen. "Jane, John ... Leslie? A Historical Method for Algorithmic Gender Prediction." *Digital Humanities Quarterly*, volume 9, number 3, 2015, <http://www.digitalhumanities.org/dhq/vol/9/3/000223/000223.html>
- Bode, Katherine. "The Equivalence of 'Close' and 'Distant' Reading; Or, Toward a New Object for Data-Rich Literary History." *Modern Language Quarterly*, volume 78, number 1, 2017, pp. 77–106, doi:10.1215/00267929-3699787.
- . *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press, 2018.
- Bogost, Ian. *Persuasive Games: The Expressive Power of Videogames*, MIT Press, 2010.
- Brewer, David. "Counting, Resonance, and Form, A Speculative Manifesto." *Eighteenth-Century Fiction*, volume 24, issue 2, 2011, pp. 161-170, doi: 10.1353/ecf.2011.0053.
- Brown, Susan, Patricia Clements, Isobel Grundy, Stan Ruecker, Jeffery Antoniuk, and Sharon Balazs. "Published Yet Never Done: The Tension Between Projection and Completion in Digital Humanities Research." *Digital Humanities Quarterly*, volume 3, number 2, 2009, digitalhumanities.org/dhq/vol/3/2/000040/000040.html
- Bruhm, Steven. "The Gothic Novel and the Negotiation of Homophobia." *The Cambridge History of Gay and Lesbian Literature*, edited by E.L. McCallum and Mikko Tuhkanen, Cambridge UP, 2014, pp. 272-87.

- Bullard, Paddy. "Digital Humanities and Electronic Resources in the Long Eighteenth Century." *Literature Compass*, volume 10, 2013, pp. 748-760.
- Buurma, Rachel Sagner, and Laura Heffernan. "Search and Replace: Josephine Miles and the Origins of Distant Reading." *Modernism / Modernity Print+*, 2 March 2016, modernismmodernity.org/forums/posts/search-and-replace. Accessed 18 April 2018.
- Cairo, Alberto. "Infographics to Explain, Data Visualizations to Explore." *The Functional Art*, 16 March 2014, web.archive.org/web/20190923005330/http://www.thefunctionalart.com/2014/03/infographics-to-reveal-visualizations.html. Accessed 22 September 2019.
- Carson, James. Review of Ann Radcliffe, *Romanticism and the Gothic*, edited by Dale Townshend and Angela Wright. *Eighteenth-Century Studies*, vol. 48, no. 1, 2014, pp. 127-129.
- Champion, Erik. "Digital humanities is text heavy, visualization light, and simulation poor." *Digital Scholarship in the Humanities*, vol. 32, supplement to issue 1, April 2017, pp. 25-32, doi:10.1093/llc/fqw053.
- Christy, Matthew, Anshul Gupta, Elizabeth Grumbach, et al. "Mass Digitization of Early Modern Texts With Optical Character Recognition." *ACM Journal on Computing and Cultural Heritage*, volume 11, number 1, article 6, December 2017, 25 pp., doi:10.1145/3075645.
- Chun, Wendy Hui Kyong. "Queerying Homophily." *Pattern Discrimination*, by Clemens Apprich, Wendy Hui Kyong Chun, Florian Cramer, and Hito Steyerl, meon press[they style it all lowercase] and U Minnesota P, 2018, pp. 59-97, doi:10.14619/1457. In *Search Of Media* series, edited by Götz Bachman, Timon Beyes, Mercedes Bunz, and Wendy Hui Kyong Chun.
- Civale, Susan. "Women's life writing and reputation: A case study of Mary Darby Robinson." *Romanticism*, vol. 24, no. 2, 2018, pp. 181-202.
- Clemens, Justin. "Aggressively middling: The Bourgeois & Distant Reading by Franco Moretti." *Sydney Review of Books*, 16 July 2013, web.archive.org/web/20200309032758/https://sydneyreviewofbooks.com/review/aggressively-middling/ Accessed 8 March 2020.
- Clery, E.J. *The Rise of Supernatural Fiction 1762-1800*. Cambridge UP, 1995.
- Christman, Paul. "The Cinema of Inadvertence." *The Hedgehog Review: Critical Reflections on Contemporary Culture*, volume 21, number 3, Fall 2019, web.archive.org/web/20191118012126/https://hedgehogreview.com/issues/eating-and-being/articles/the-cinema-of-inadvertence-or-why-i-like-bad-movies
- Cronin, Richard. *The Politics of Romantic Poetry: In Search of the Pure Commonwealth*. Palgrave Macmillan, 2000.

- Cross, Ashley. "From Lyrical Ballads to Lyrical Tales: Mary Robinson's Reputation and the Problem of Literary Debt." *Studies in Romanticism*, vol. 40, no. 4, 2001, pp. 571–605, doi:10.2307/25601532.
- . *Mary Robinson and the Genesis of Romanticism: Literary Dialogues and Debts, 1784–1821*. Routledge, 2016.
- Cohen, Margaret. *The Sentimental Education of the Novel*. Princeton UP, 1999.
- Coker, Cait, and Kate Ozment. "Building the Women in Book History Bibliography, or Digital Enumerative Bibliography as Preservation of Feminist Labor." *Digital Humanities Quarterly*, volume 13, number 3, 2019, www.digitalhumanities.org/dhq/vol/13/3/000428/000428.html.
- Cooke, Richard. "Wikipedia Is the Last Best Place on the Internet." *Wired*, 17 February 2020, web.archive.org/web/20200227222807/https://www.wired.com/story/wikipedia-online-encyclopedia-best-place-internet/
- Crump, M. J. "Short Title Catalogue On-Line." *Information Development*, vol. 2, no. 2, April 1986, pp. 105–107, doi:10.1177/026666698600200208.
- Curran, Stuart. "Charlotte Smith: Intertextualities." *Charlotte Smith in British Romanticism*, edited by Jacqueline Labbe. Pickering & Chatto, 2008, pp. 175–188.
- Dawkins, Richard. "Memes: the new replicators." 1976. *The Selfish Gene*, Oxford UP, 1989, pp. 189–201.
- Dayal, Samir. Review of *Cultural Capital: The Problem of Literary Canon Formation*, by John Guillory. *MELUS*, volume 21, number 2, June 1996, pp. 165–168, doi:10.2307/467957.
- Denham, Alison. "Making Sorrow Sweet: Emotion and Empathy in the Experience of Fiction." *Affect and Literature*, ed. Alex Houen, Cambridge UP, 2020, pp. 190–210, doi:10.1017/9781108339339.011.
- Drucker, Johanna. *Graphesis: Visual Forms of Knowledge Production*. Harvard UP, 2014.
- . "Humanities Approaches to Graphical Display." *Digital Humanities Quarterly*, vol. 5, no. 1, 2011, www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.
- Duckling, Louise. "'Tell My Name to Distant Ages': The Literary Fate of Charlotte Smith." *Charlotte Smith in British Romanticism*, edited by Jacqueline Labbe. Pickering & Chatto, 2008, pp. 203–217.
- "ECCO-TCP: Eighteenth Century Collections Online." Text Creation Partnership, www.textcreationpartnership.org/tcp-ecco. Accessed 20 June 2019.
- Emre, Merve. "Public Thinker: Leah Price on Books, Book Tech, and Book Tattoos." Interview with Leah Price. *Public Books*, 26 Sept 2019, web.archive.org/web/20190928023628/https://www.publicbooks.org/public-thinker-leah-price-on-books-book-tech-and-book-tattoos. Accessed 27 Sept 2019.

- Ezell, Margaret J. M. "Big Books, Big Data, and Reading Literary Histories." *Eighteenth-Century Life*, volume 41, issue 3, September 2017, pp. 3-19. Project MUSE, muse.jhu.edu/article/667488.
- Facer, Ruth. "Ann Radcliffe (1764-1823)." *Women Writer Biographies*, Chawton House, chawtonhouse.org/the-library/library-collections/womens-writing-in-english/women-writer-biographies/.
- Falkovich, Jacob. "100 Ways to Live Better." *Put A Number On It!*, 30 December 2019, web.archive.org/save/https://putanumonit.com/2019/12/30/100-ways-to-live-better/.
- Farmer, Alan B., and Zachary Lesser. "Structures of Popularity in the Early Modern Book Trade." *Shakespeare Quarterly*, vol. 56, no. 2, 2005, pp. 206-213. JSTOR, www.jstor.org/stable/3844307.
- . "The Popularity of Playbooks Revisited." *Shakespeare Quarterly*, vol. 56, no. 1, 2005, pp. 1-32. JSTOR, www.jstor.org/stable/3844024.
- Felski, Rita. "Everyday Aesthetics." *the minnesota review* [they style it all lowercase], volume 71-72, 2009, pp. 171-179.
- . *The Limits of Critique*. U Chicago P, 2015.
- Finley, Klint. "The Internet Archive Is Making Wikipedia More Reliable." *Wired*, 3 November 2019, web.archive.org/web/20200227222720/https://www.wired.com/story/internet-archive-wikipedia-more-reliable/.
- Fischer-Starcke, Bettina. *Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries*. Continuum, 2010.
- Flores, Pepe. "Is Wikipedia the largest-ever digital humanities project? Exploring an emerging relationship." *Wikimedia Foundation blog*, 17 April 2016, web.archive.org/web/20200428033423/https://blog.wikimedia.org/2016/08/17/wikipedia-largest-digital-humanities-project/.
- Forster, Chris. "A Walk Through the Metadata: Gender in the HathiTrust Dataset." 8 Sept. 2015, cforster.com/2015/09/gender-in-hathitrust-dataset. Accessed 3 Sept. 2019.
- Fowers, Alyssa. "Profiling protest data (or, what I did on my summer vacation)." *Data and Dragons*, 10 Sept 2019, dataanddragons.wordpress.com/2019/09/10/profiling-protests-or-what-i-did-on-my-summer-vacation. Accessed 10 Sept 2019.
- Frank, Marcie. "Melodrama and the Politics of Literary Form in Elizabeth Inchbald's Works." *Eighteenth-Century Fiction*, volume 27, number 3-4, Spring-Summer 2015, pp. 707-730.
- Freedgood, Elaine. "Reading Things." *The Ideas in Things: Fugitive Meaning in the Victorian Novel*, U Chicago P, 2006.
- Frow, John. *Genre*. Routledge, 2015.

- Fry, Carrol L. Charlotte Smith. Twayne's English Authors Series, edited by Herbert Sussman, Twayne, 1996.
- Gadd, Ian. "The Use and Misuse of Early English Books Online." *Literature Compass*, volume 6, issue 3, 2009, pp. 680-692, doi:10.1111/j.1741-4113.2009.00632.x.
- Gale. "Eighteenth Century Collections Online."
web.archive.org/web/20200324195501/https://www.gale.com/primary-sources/eighteenth-century-collections-online Accessed 24 March, 2020.
- Gamer, Michael. *Romanticism and the Gothic: Genre, Reception, and Canon Formation*. Cambridge UP, 2000.
- Gamer, Michael, and Terry F. Robinson. "Mary Robinson and the Dramatic Art of the Comeback." *Studies in Romanticism*, vol. 48, no. 2, 2009, pp. 219-56. JSTOR, www.jstor.org/stable/25602191.
- Garnai, Amy. *Revolutionary Imaginings in the 1790s: Charlotte Smith, Mary Robinson, Elizabeth Inchbald*. Palgrave Macmillan, 2009.
- Garside, Peter. "The English Novel in the Romantic Era: Consolidation and Dispersal." *The English Novel 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles*, edited by Peter Garside, James Raven, and Rainer Schöwerling, vol. 2: 1800-1829, edited by Peter Garside and Rainer Schöwerling with Christopher Skelton-Foord and Karin Wünsche. Oxford UP, 2000.
- Garside, Peter, James Raven, and Rainer Schöwerling, editors. *The English Novel 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles*, Oxford UP, 2000. 2 vols.
- . General Introduction. *The English Novel 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles*, edited by Peter Garside, James Raven, and Rainer Schöwerling, Oxford UP, 2000. 2 vols.
- Gavin, Michael. "Historical Text Networks: The Sociology of Early English Criticism." *Eighteenth-Century Studies*, volume 50, number 1, 2016, pp. 53-80.
- Gilbert, Geoff. "The Durability of Affect and the Ageing of Gay Male Queer Theory." *Affect and Literature*, ed. Alex Houen, Cambridge UP, 2020, pp. 133-158, doi: 10.1017/9781108339339.008.
- Goldie, Mark and Robert Wokler, editors. *The Cambridge history of eighteenth-century political thought*. Cambridge UP, 2006.
- Gonda, Caroline. "Review of Heteronormativity in Eighteenth-Century Literature and Culture, ed. by Ana de Freitas Boe and Abby Coykendall." *Eighteenth Century Studies*, volume 49, issue 3, 2016, pp. 427-428.

- Google Books. "Google Books History." Site as of 25 March 2020,
web.archive.org/web/20200326031915/https://books.google.com/googlebooks/about/history.html Accessed 25 March 2020.
- . "Google Books History." Site as of 6 February 2016,
<https://web.archive.org/web/20160206043510/http://books.google.com/googlebooks/about/history.html>. Accessed 25 March 2020.
- Gorak, Jan. Review of *Cultural Capital: The Problem of Literary Canon Formation*, by John Guillory. *Modern Philology*, volume 94, number 2, Nov 1996, pp. 286-290. JSTOR, www.jstor.org/stable/437977.
- Grafton, Anthony and Glenn W. Most. "How to do things with texts: An introduction." *Canonical Texts and Scholarly Practices: A Global Comparative Approach*. Cambridge UP, 2016, pp. 1-13. doi:10.1017/CBO9781316226728.001.
- Gregg, Stephen H. "Finding ECCO-TCP texts." *Manicule: Thoughts on the Eighteenth Century, Daniel Defoe, and Digital Humanities*, Wordpress, 16 Aug. 2017, shgregg.com/2017/08/16/finding-ecco-tcp-texts. Accessed 20 June 2019.
- Guillory, John. *Cultural Capital: The Problem of Literary Canon Formation*. U Chicago P, 1993.
- Hammond, Adam. *Literature in the Digital Age: An Introduction*. Cambridge UP, 2016.
- Hane, Paula. "Project Gutenberg Progresses." *Information Today*, volume 21, number 5, May 2004,
web.archive.org/web/20200325195437/http://www.infotoday.com/it/may04/hanel.shtml. Accessed 25 March 2020.
- Hart, Michael. "The History and Philosophy of Project Gutenberg." Project Gutenberg, August 1992.
web.archive.org/web/20200312224522/https://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart Accessed 12 March 2020.
- HathiTrust. "Getting Content Into HathiTrust." HathiTrust,
web.archive.org/web/20190915204356/https://www.hathitrust.org/ingest. Accessed 15 Sept 2019.
- . "Governance." HathiTrust,
web.archive.org/web/20200325212223/https://www.hathitrust.org/governance. Accessed 25 March 2020.
- . "Help - Copyright." HathiTrust,
web.archive.org/web/20200325212528/https://www.hathitrust.org/help_copyright%2RestrictedAccess. Accessed 25 March 2020.

- . "Major Library Partners Launch HathiTrust Shared Digital Repository." HathiTrust, 13 October 2008,
web.archive.org/web/20200325213628/https://www.hathitrust.org/press_10-13-2008.
 Accessed 25 March 2020.
- . "Member Community." HathiTrust,
web.archive.org/web/20190915204844/https://www.hathitrust.org/community.
 Accessed 15 Sept 2019.
- . "Our Digital Library." HathiTrust,
web.archive.org/web/20190915204611/https://www.hathitrust.org/digital_library.
 Accessed 15 Sept 2019.
- . "Our Membership." HathiTrust,
web.archive.org/web/20190915204720/https://www.hathitrust.org/partnership.
 Accessed 15 Sept 2019.
- Heisel, Andrew. "Hannah More's Art of Reduction." *Eighteenth-Century Fiction*, volume 25, number 3, Spring 2013, pp. 557-588.
- Hosch, William L. "Project Gutenberg." *Encyclopedia Britannica*, 17 August 2017,
web.archive.org/web/20200325190246/https://www.britannica.com/topic/Project-Gutenberg. Accessed 25 March 2020.
- Hunt, Bishop C. "Wordsworth and Charlotte Smith." *The Wordsworth Circle*, vol. 1, no. 3, 1970, pp. 85. ProQuest, ProQuest Document ID 1300171026.
- IPUMS NAPP. "What is NAPP?" North Atlantic Population Project,
web.archive.org/web/20200209014205/https://www.nappdata.org/napp/intro.shtml.
 Accessed February 8, 2020.
- Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History*. U Illinois P, 2013.
- Juxta Commons. "A User Guide to Juxta Commons."
web.archive.org/web/20200227014953/http://juxtacommons.org/guide.
- Karels, Liene. "HathiTrust adds new members, goes global." *Montage: Arts + Creativity*, University of Michigan, November 2010,
web.archive.org/web/20140302084528/http://www.montage.umich.edu/2010/11/hathitrust-adds-new-members-goes-global. Accessed 25 March 2020.
- Karian, Stephen. "The Limitations and Possibilities of the ESTC." *The Age of Johnson*, vol. 21, 2011, pp. 283-297. ProQuest, ProQuest document ID 1689625001.
- King, Kathryn R. "Introduction: Hans Turley, Queer Studies, and the Open-Hatched Eighteenth Century." *The Eighteenth Century*, volume 53, number 3, 2012, pp. 265-272. JSTOR, www.jstor.org/stable/23365012.

- Klein, Lauren. "Distant Reading After Moretti." ["the text of a talk delivered at the 2018 MLA Annual Convention for a panel, "Varieties of Digital Humanities," organized by Alison Booth and Miriam Posner. Marisa Parham, Alan Liu, and Ted Underwood were the other speakers. (Howard Ramsby was also scheduled to present, but he was unable to attend because of the blizzard.)"] *Arcade: Literature, the Humanities, & the World*, 2018, web.archive.org/save/https://arcade.stanford.edu/blogs/distant-reading-after-moretti. Accessed 19 September 2019.
- Klein, Ula, and Emily MN Kugler. "Eighteenth-Century Camp Introduction." *ABO: Interactive Journal for Women in the Arts, 1640-1830*, volume 9, issue 1, 2019, pp. 1-12. doi:10.5038/2157-7129.9.1.1180
- Korshin, Paul J. Review of *Bibliography, Machine Readable Cataloguing, and the ESTC. A Summary History of the Eighteenth Century Short Title Catalogue. Working Methods. Cataloguing Rules. A Catalogue of the Works of Alexander Pope Printed Between 1711 and 1800 in the British Library*, by R. C. Alston and M. C. Jannetta. *Eighteenth-Century Studies*, volume 12, number 2, 1978, pp. 209–212. JSTOR, www.jstor.org/stable/2738046.
- Labbe, Jacqueline. "Introduction." *Charlotte Smith in British Romanticism*, edited by Jacqueline Labbe. Pickering & Chatto, 2008, pp. 1-11.
- . "Selling One's Sorrows: Charlotte Smith, Mary Robinson, and the Marketing of Poetry." *The Wordsworth Circle*, volume 25, number 2, 1994, pp. 68-71. ProQuest, ProQuest Document ID 1300173031.
- LaGuardia, Cheryl. Review of *Eighteenth Century Collections Online*. *Library Journal*, May 2004, pp. 123-124.
- Lahti, Leo, Niko Ilomäki, and Mikko Tolonen. "A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800." *LIBER Quarterly*, volume 25, number 2, 2015, pp. 87–31, doi:10.18352/lq.10112.
- Lebert, Marie. "Project Gutenberg (1971-2008)." Project Gutenberg, May 2008, www.gutenberg.org/cache/epub/27045/pg27045-images.html Accessed 12 March 2020
- Liberman, Mark. "The 'dance of the p's and b's': truth or noise?" *Language Log*, 26 Jan 2012, web.archive.org/web/20191105001115/https://languagelog.ldc.upenn.edu/nll/?p=3730. Accessed 4 Nov 2019.
- Love, Heather. "Close but Not Deep: Literary Ethics and the Descriptive Turn." *New Literary History*, volume 41, issue 2, 2010, pp. 371–391.
- . "Close Reading and Thin Description." *Public Culture*, volume 25, number 3 (71), 2013, pp. 401-434, doi:10.1215/08992363-2144688.

- "MARC 21 Format for Authority Data." Cataloger's Reference Shelf, The Library Corporation, www.itsmarc.com/crs/mergedProjects/helpauth/helpauth/Contents.htm.
- Marche, Stephen. "Literature Is not Data: Against Digital Humanities." *Los Angeles Review of Books*, 28 Oct. 2012.
web.archive.org/web/20191022060530/https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/
- Mark Ockerbloom, Mary. "Mary Darby Robinson (1758-1800)." *A Celebration of Women Writers*, digital.library.upenn.edu/women/robinson/biography.html. Accessed 07 June 2019.
- Mauri, M., T. Elli, G. Caviglia, G. Ubaldi, and M. Azzi. (2017). "RAWGraphs: A Visualisation Platform to Create Open Outputs." *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*, ACM, 2017, p. 28:1–28:5, doi:10.1145/3125571.3125585.
- McCarty, Willard. "Knowing: Modeling in Literary Studies." *A Companion to Digital Literary Studies*, edited by Susan Schreibman and Ray Siemens, Blackwell, 2008, www.digitalhumanities.org/companionDLS/.
- McLavery, James. "Poems in Print." *The Oxford Handbook of British Poetry, 1660-1800*, edited by Jack Lynch. Oxford UP, 2016, pp. 40-54, doi:10.1093/oxfordhb/9780199600809.013.3. Oxford Handbooks Online.
- McLeod, Dayna, Jasmine Rault, and T.L. Cowan. "Speculative Praxis Towards a Queer Feminist Digital Archive: A Collaborative Research-Creation Project." *Ada: A Journal of Gender, New Media, & Technology*, issue 5, 2014, web.archive.org/web/20190318202624/https://adanewmedia.org/2014/07/issue5-cowanetal/
- McLeod, Deborah Anne. The Minerva Press. PhD dissertation, U of Alberta, 1997, doi:10.7939/R33J39C22.
- McKitterick, David. "Obituary: Katharine F. Pantzer, 1930-2005." *The Library: The Transactions of the Bibliographical Society*, vol. 7, no. 1, March 2006, pp. 87-89. Project MUSE, muse.jhu.edu/article/203028.
- Mee, John. *Print, Publicity, and Popular Radicalism in the 1790s: The Laurel of Liberty*. Cambridge UP, 2016.
- Moretti, Franco. *The Bourgeois: Between History and Literature*. Verso, 2013.
- . "Conjectures on World Literature." *New Left Review*, volume 1, issue 1, 2000, pp. 54 - 67.
- . *Distant Reading*. Verso, 2013.
- Mullen, Lincoln. "gender: Predict Gender from Names Using Historical Data." R package version 0.5.2. GitHub, <https://github.com/ropensci/gender>
- Mullen, Lincoln, Cameron Blevins, and Ben Schmidt. "Package 'gender.'" November 9, 2019.

- Murphie, Andrew. "The Digital's Amodal Affect." *Affect and Literature*, ed. Alex Houen, Cambridge UP, 2020, pp. 390-407, doi: 10.1017/9781108339339.022.
- Murphy, Peter. *Poetry as an occupation and an art in Britain, 1760-1830*. Cambridge UP, 1993.
- Nicolazzo, Sarah. "Reading Clarissa's "Conditional Liking": A Queer Philology." *Modern Philology*, volume 112, issue 1, 2014, pp. 205-225. JSTOR, www.jstor.org/stable/10.1086/676008.
- Noble, Safiya. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.
- Norman, Jeremy. "The English Short Title Catalogue (ESTC) is Conceived: 6/1976." Jeremy Norman's History of Information, www.historyofinformation.com/detail.php?entryid=2915. Accessed 26 June 2019.
- OpenRefine. Version 3.1, Nov. 29, 2018, openrefine.org.
- O'Dair, Sharon. Review of *Cultural Capital: The Problem of Literary Canon Formation*, by John Guillory. *South Atlantic Review*, volume 59, number 2, May 1994, pp. 130-132. JSTOR, www.jstor.org/stable/3200802.
- O'Quinn, Daniel. "Half-History, or The Function of Cato at the Present Time." *Georgian Theatre in an Information Age: Media, Performance, Sociability*, special issue of *Eighteenth-Century Fiction*, vol. 27, no. 3-4, 2015, pp. 479-507. Project Muse, muse.jhu.edu/article/584623.
- Potter, Franz J. *The History of Gothic Publishing, 1800-1835: Exhuming the Trade*. Palgrave Macmillan, 2005.
- Price, Leah. *The Anthology and the Rise of the Novel: From Richardson to George Eliot*. Cambridge UP, 2000, doi:10.1017/CBO9780511484445.
- Prior, Karen Swallow. "Hannah More." *The Literary Encyclopedia*, volume 1.2.1.06: *English Writing and Culture of the Romantic Period, 1789-1837*, edited by Daniel Cook and Daniel Robinson, 16 Dec. 2008, www.litencyc.com. Accessed 29 June 2019.
- Project Gutenberg. "Credits." 7 June 2006, web.archive.org/web/20200325192439/https://www.gutenberg.org/wiki/Gutenberg:Credits. Accessed 25 March 2020.
- . "The CD and DVD Project." 19 August 2017, web.archive.org/web/20200325194042/https://www.gutenberg.org/wiki/Gutenberg:The_CD_and_DVD_Project. Accessed 25 March 2020.
- . "Partners, Affiliates and Resources." 24 January 2019, web.archive.org/web/20200325193013/https://www.gutenberg.org/wiki/Gutenberg:Partners,_Affiliates_and_Resources. Accessed 25 March 2020.

- Punter, David. Review of *Cultural Capital: The Problem of Literary Canon Formation*, by John Guillory. *Non-Standard Englishes and the New Media*, special issue of *The Yearbook of English Studies*, volume 25, 1995, pp. 229-230. JSTOR, www.jstor.org/stable/3508832.
- Raven, James. "Historical Introduction: The Novel Comes of Age." *The English Novel 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles*, edited by Peter Garside, James Raven, and Rainer Schöwerling, vol. 1: 1770-1799, edited by James Raven and Antonia Forster with Steven Bending, Oxford UP, 2000.
- Readings, Bill. Review of *Cultural Capital: The Problem of Literary Canon Formation*, by John Guillory. *Modern Language Quarterly*, volume 55, number 3, Sept 1994, pp. 321-326, doi:10.1215/00267929-55-3-321.
- @RealSaavedra (Ryan Saavedra). "Socialist Rep. Alexandria Ocasio-Cortez (D-NY) claims that algorithms, which are driven by math, are racist." Twitter, 22 Jan. 2019, 3:27 AM, web.archive.org/save/https://twitter.com/realsaavedra/status/1087627739861897216?lang=en.
- Reason. "How to Politely Download All English Language Text Format Files from Project Gutenberg." *Ex Ratione*, 1 November 2014, web/20200506014409/https://www.exratione.com/2014/11/how-to-politely-download-all-english-language-text-format-files-from-project-gutenberg/
- Reinert, Thomas. Review of *Cultural Capital: The Problem of Literary Canon Formation*, by John Guillory. *Modern Fiction Studies*, volume 42, number 1, 1996, pp. 221-224. ProQuest, ProQuest Document ID 2152910014.
- Rigby, Mair. "Uncanny recognition: Queer theory's debt to the Gothic." *Gothic Studies*, vol. 11, no. 1, 2009, pp. 46-57.
- Roberts, Bethan. "Charlotte Smith: Elegiac Sonnets, and Other Essays." Edited by Daniel Robinson. *The Literary Encyclopedia*, volume 1.2.1.06: English Writing and Culture of the Romantic Period, 1789-1837, edited by Daniel Cook and Daniel Robinson, 02 January 2014, www.litencyc.com. Accessed 05 June 2019.
- Rogers, Deborah D. *Ann Radcliffe: A Bio-Bibliography*. *Bio-Bibliographies in World Literature*, number 4. Greenwood Press, 1996.
- Rosenberg, Scott. "How Google Book Search Got Lost." *Wired*, 11 April 2017, web.archive.org/web/20200326035223/https://www.wired.com/2017/04/how-google-book-search-got-lost/. Accessed 25 March 2020.
- Runge, Laura. "Mary Darby Robinson (1758?-1800) - Bibliography." chuma.cas.usf.edu/~runge/MRobinson.htm. Accessed 07 June 2019.

- Sandvoss, Cornel. "The Death of the Reader: Literary Theory and the Study of Texts in Popular Culture." *Fandom: Identities and Communities in a Mediated World*, edited by Jonathan Gray, C. Sandvoss, and C. Lee Harrington. NYU Press.
- Seaver, Nick. "Bastard Algebra." *Data, Now Bigger and Better!*, edited by Tom Boellstorff and Bill Maurer. Prickly Paradigm, 2015.
- Sedgwick, Eve Kosofsky. "Paranoid Reading and Reparative Reading: Or, You're So Paranoid, You Probably Think This Essay is About You." *Touching Feeling*, Duke UP, 2003, pp. 123-151.
- Shaw, Zed. *Learn Python the Hard Way*, Addison-Wesley Professional, 2013.
- Shirky, Clay. "Why Abundance is Good: A Reply to Nick Carr." *Encyclopædia Britannica Blog*, 17 July 2008, blogs.britannica.com/2008/07/why-abundance-is-good-a-reply-to-nick-carr. Accessed 10 Sept 2019.
- Shiroma, Yumi Dineen. "Conjectures on World Literature, Revisited." June 3, 2018, [web.archive.org/web/20200309032430/http://yumidineenshiroma.org/blog/conjectures-on-world-literature-revisited/](http://yumidineenshiroma.org/blog/conjectures-on-world-literature-revisited/) Accessed March 1, 2020.
- Somers, James. "Torching the Modern-Day Library of Alexandria." *The Atlantic*, 20 April 2017, [web.archive.org/web/20200326035404/https://www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/](https://www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/). Accessed 25 March 2020.
- Spedding, Patrick. "'The New Machine': Discovering the Limits of ECCO." *Eighteenth-Century Studies*, volume 44, issue 4, 2011, pp. 437-453.
- St. Clair, William. *The Reading Nation in the Romantic Period*. Cambridge UP, 2007.
- Stanton, Judith Phillips. "Recovering Charlotte Smith's Letters: A History, With Lessons." *Charlotte Smith in British Romanticism*, edited by Jacqueline Labbe. Pickering & Chatto, 2008, pp. 159-173.
- Stott, Anne. "Hannah More Chronology." *The Victorian Web*, 2002, www.victorianweb.org/authors/more/chron.html.
- Suarez, Michael F. "Towards a bibliometric analysis of the surviving record, 1701–1800." *The Cambridge History of the Book in Britain, Volume 5: 1695–1830*, edited by Michael F. Suarez and Michael L. Turner, Cambridge UP, 2009, pp. 39-65.
- Syme, Holger Schott. "Imaginary Targets." *Los Angeles Review of Books*, 5 Nov 2012. [web.archive.org/web/20191023050529/https://lareviewofbooks.org/article/in-defense-of-data-responses-to-stephen-marches-literature-is-not-data](https://lareviewofbooks.org/article/in-defense-of-data-responses-to-stephen-marches-literature-is-not-data).
- Tabor, Stephen. "ESTC and the Bibliographical Community." *The Library: The Transactions of the Bibliographical Society*, vol. 8 no. 4, 2007, pp. 367-386. Project MUSE, muse.jhu.edu/article/230381. (If I add new Tabor citations, go back and clarify which ones I'm already quoting)

Taylor, George. *The French Revolution and the London Stage, 1789–1805*. Cambridge UP, 2001.

TCP Executive Board. “Meeting Minutes 2001-01-12.” [wayback.archive-it.org/5871/20190806191909/http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2001-01-12/](http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2001-01-12/) Accessed 10 March 2020.

—. “Meeting Minutes 2003-10-22.” [wayback.archive-it.org/5871/20190806191856/http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2003-10-22/](http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2003-10-22/) Accessed 23 March 2020.

—. “Meeting Minutes 2004-10-21.” [wayback.archive-it.org/5871/20190806191851/http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2004-10-21/](http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2004-10-21/) Accessed 23 March 2020.

—. “Meeting Minutes 2005-10-20.” [https://wayback.archive-it.org/5871/20190806191847/http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2005-10-20/](http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2005-10-20/) Accessed 23 March 2020.

—. “Meeting Minutes 2006-09-16.” [https://wayback.archive-it.org/5871/20190806191843/http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2006-09-16/](http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2006-09-16/) Accessed 23 March 2020.

—. “Meeting Minutes 2007-10-30.” [https://wayback.archive-it.org/5871/20190806191838/http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2007-10-30/](http://www.textcreationpartnership.org/tcp-board-meeting-minutes-2007-10-30/) Accessed 23 March 2020.

Text Creation Partnership (TCP). “About the partnership.” [web.archive.org/web/20200312203119/https://textcreationpartnership.org/about-the-tcp/](https://textcreationpartnership.org/about-the-tcp/). Accessed 12 March 2020.

—. “Eighteenth Century Collections Online (ECCO) TCP.” [web.archive.org/web/20200312212808/https://textcreationpartnership.org/tcp-texts/ecco-tcp-eighteenth-century-collections-online/](https://textcreationpartnership.org/tcp-texts/ecco-tcp-eighteenth-century-collections-online/). Accessed 12 March 2020.

—. “Early English Books Online (EEBO) TCP.” [web.archive.org/web/20200312212215/https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/](https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/). Accessed 12 March 2020.

—. “FAQ.” [web.archive.org/web/20200311042209/https://textcreationpartnership.org/faq/](https://textcreationpartnership.org/faq/). Accessed 10 March 2020.

—. “Our scholarly partners.” [web.archive.org/web/20200311042125/https://textcreationpartnership.org/about-the-tcp/about-partner-libraries/](https://textcreationpartnership.org/about-the-tcp/about-partner-libraries/). Accessed 10 March 2020.

—. “Welcome.” [web.archive.org/web/20200311042006/https://textcreationpartnership.org/](https://textcreationpartnership.org/). Accessed 10 March 2020.

- Thaventhiran, Helen. "Feelings under the Microscope: New Critical Affect." *Affect and Literature*, ed. Alex Houen, Cambridge UP, 2020, pp. 83-99, doi:10.1017/9781108339339.005.
- Townshend, Dale and Angela Wright. "Gothic and Romantic engagements: The critical reception of Ann Radcliffe, 1789–1850." *Ann Radcliffe, Romanticism and the Gothic*, Cambridge University Press, 2014.
- Tufte, Edward. *The Visual Display of Quantitative Information*. 2nd ed. Graphics Press, 2001.
- Underwood, Ted. "A dataset for distant-reading literature in English, 1700-1922." *The Stone and the Shell*, 7 August 2015, <https://web.archive.org/web/20200207044631/https://tedunderwood.com/2015/08/07/a-dataset-for-distant-reading-literature-in-english-1700-1922/>. Accessed 6 Feb 2020.
- . "A Genealogy of Distant Reading." *Digital Humanities Quarterly*, volume 11, issue 2, 2017, www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.
- . *Distant Horizons: Digital Evidence and Literary Change*. U Chicago P, 2019.
- . "Do humanists get their ideas from anything at all?" *The Stone and the Shell*, 24 Jan 2012, web.archive.org/web/20191105004437/https://tedunderwood.com/2012/01/24/discovery-and-hypothesis-testing. Accessed 4 Nov 2019.
- . *Why Literary Periods Mattered: Historical Contrast and the Prestige of Literary Studies*. Stanford UP, 2013.
- Underwood, Ted and David Bamman. "The instability of gender." *The Stone and the Shell*, 9 January 2016, <https://web.archive.org/web/20200207044340/https://tedunderwood.com/2016/01/09/the-instability-of-gender/>. Accessed 6 Feb 2020.
- . "Preregistered Hypotheses for Evaluating Models of Literary Character," June 2014, hdl.handle.net/2142/49936.
- Vander Meulen, David. "ESTC as Foundational and Always Developing." *The Age of Johnson*, vol. 21, 2001, pp. 263-282.
- Vara, Vauhini. "Project Gutenberg Fears No Google." *The Wall Street Journal Online*, 10 December 2005, <https://www.wsj.com/articles/SB113415403113218620>. NOT ACCESSED IN FULL
- Walker, William. "Aroused Yet Thoughtful: Readers in Eighteenth-Century Britain." *Review of Excitable Imaginations: Eroticism and Reading in Britain, 1660-1760*, by Kathleen Lubey. *Eighteenth Century Life*, volume 39, number 2, April 2015, pp. 87-91.
- Watt, Ian. *The Rise of the Novel: Studies in Defoe, Richardson and Fielding*. 1957. U of California P, 2001.

- Wheeles D., Jensen K. (2013). Juxta Commons. In Proceedings of the Digital Humanities 2013. University of Nebraska-Lincoln, 17 July 2013. <http://dh2013.unl.edu/abstracts/ab-142.html>.
- Wilkins, Matt. "Literary Attention Lag." Work Product, 13 January 2015, web.archive.org/web/20200211050232/https://mattwilkens.com/2015/01/13/literary-attention-lag/
- Zimmerman, Sarah M. "Smith [née Turner], Charlotte (1749–1806), poet and novelist." Oxford Dictionary of National Biography, Oxford University Press, 4 Oct. 2007, doi: 10.1093/ref:odnb/25790. Accessed 13 July 2019.
- Zwicker, Steven N. "Is There Such a Thing as Restoration Literature?" *Huntington Library Quarterly*, vol. 69, no. 3, 2006, pp. 425–450, doi:10.1525/hlq.2006.69.3.425.