Canonical Corpora of English Literature, 1789-99

Dissertation chapters submitted for consideration for the Jackman Humanities Institute Chancellor Jackman Graduate Student Fellowships in the Humanities

Approved by Alex Gillespie, supervisor

Lawrence Evalyn
March 2, 2020

Chapter One: Introduction

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 inflammatory 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. These claims demonstrate that a poorly-formed question will produce a useless and stupid answer even (or perhaps especially) if computation is used to 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. It is also an old, nearly an old-fashioned question. The current moment of self-reflection in the field of Digital Humanities,

[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."

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 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.

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, the English Short Title Catalogue provides basic bibliographic data for nearly 52,000 titles, but the Eighteenth Century Collections Online Text Creation Partnership 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: do the most invested-in resources underrepresent women? 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 smaller databases. 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 selection. To ground my analysis in specifics, I take Charlotte Smith, Mary Robinson, Hannah More, and Ann Radcliffe as case study authors. All four authors have long histories of contentious reception, rooted in debates about seriousness, popularity, and women's writing. I revisit them to see how their careers might be interpreted through a new lens. I do so,

[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.

in part, to challenge the contrast drawn between 'popular' and 'serious' writing, especially in the historical evaluation of women's writing as literary.

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 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.

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). 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 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

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

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.

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

[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.

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 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.

2. Frameworks

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. However, these obvious critiques also require rejecting many generations of sincere work by my 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 resources to be recuperated. My touchstones are two descriptions from Sedgwick's original chapter:

> 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)

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

[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.

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.

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:

> 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 Thomasticus* 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

[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. (Buurma and Heffernan)

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 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

[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.

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. In other words, in a racist world, algorithms *are* racist — and sexist, and duplicative of all other systemic inequities.

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.

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' pre-exist 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 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)

---

[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.

[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.

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 dangerous, and potentially averted disaster.[14] Tufte's six principles of design 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. His preceding book, *The Truthful Art*, addresses the creators of good-faith infographics with insights into visual manipulation. 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". 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.

[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.

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

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.[15] 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 was the epoch of belief, it was the epoch of incredulity. Charles Dickens' now-famous superlatives capture the tension often seen by scholars between 'Enlightenment' modes of writing and 'Romantic' or 'Gothic' modes.

Scholarship on eighteenth-century 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' monograph, too, seem to treat Scott's intensely serious popular romances as the teleological end of the late eighteenth century birth development within the novel. These works follow a pattern established from the beginning

[15] 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.

with Kiely and 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 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 examines Wordsworth and the 'winners' of the struggle for cultural capital: I, like Clery, am more interested in the 'losers.' Accordingly, I attend to much that is *not* literature, in order to better understand why it is not.

To navigate the 1790s, I turn to four authors whose careers and works usefully focalize my core questions of genre, publics, and the status of literature. These authors are Hannah More, Charlotte Smith, Mary Robinson, and Ann Radcliffe. All four authors were highly productive in the 1790s, and all four had complex and contested literary legacies after the 1790s. As literary scholars re-assess ideas about literary seriousness, popularity, and women's writing, our assessment of these authors has shifted as well. By examining their bibliographies with computational methods, I again ask how they might look different if we look at them a different way. In the following chapters of this dissertation, I will ask whether and how contemporary digital archives make these authors visible. To introduce each of those chapters, I provide here a brief overview of each author's major works, general biography, and critical reception.

4.1. Charlotte Smith

Chapter 2 of this dissertation examines the works of Charlotte Smith across several digital corpora. 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, 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 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 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 wrangling 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 Leman. 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 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 considered as 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 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 *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.

The specific experimentation undertaken in chapter two tests the basic assumptions and methods of my project. 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.

4.2. Mary Robinson

Chapter three of this dissertation examines the works of Mary Robinson across several digital corpora. Mary Robinson is selected as a contemporary of Charlotte Smith's with remarkably similar patterns of publication, but remarkably different reception. Mary Robinson's first literary foray was the volume *Poems by Mrs. Robinson*, published by C. Parker in 1775, shortly followed by the poem *Elegiac verses to a young lady on the death of her brother*. "Though *Poems* received little critical support, and made little money, Mary Robinson continued to write" (Mark Ockerbloom), but she would begin her career as an actress before this writing saw publication. Robinson's theatrical debut as Juliet in December 1776 led to an increasingly acclaimed and prolific acting career. In 1777 she published a volume of works written earlier, *Captivity, A Poem: and Celadon and Lydia, A Tale,* and in 1778 she wrote and starred in her own musical farce, *The Lucky Escape,* from which she published *The Songs, Chorusses, etc.* In 1780,

Robinson retired from the stage. The next volume under her name did not appear until 1791, but the intervening decade was not empty of literary activity. From 1784 to its publication in 1787, Robinson collaborated with Banastre Tarleton on the composition and revision of his *History of the Campaigns of 1780 and 1781*. In 1788, Robinson began the writing which would make her career as an author. She contributed poems to *The World* and *The Oracle* under the pseudonyms "Laura" and "Laura Maria" which responded to popular poems by a coterie of writers now called the Della Cruscans, who addressed each other in periodicals with Miltonic, Italianate, and political poetry. Robinson's poems successfully inserted herself into their conversation, prompting replies from the other poets. This dissertation's decade of interest, 1789-99, captures Robinson's most prolific period. Her 1790 poem *Ainsi va le Monde*, published under the pseudonym Laura Maria, was her breakout Della Cruscan poem. The success of her Della Cruscan poetry led to the publication in 1791 of another *Poems by Mrs. Robinson*, this time by J. Bell, as well as *The Beauties of Mrs. Robinson*, another collection of poetry. She also produced a pamphlet that year, *Impartial Reflections on the Queen of France,* under the name "a friend to humanity". The next year, 1792, saw the arrival of a successful novel, *Vancenza, or the Dangers of Credulity*, which went through three editions in that year, and two elegiac poems. 1793 saw three more poems, two under her own name and one under the pseudonym "Horace Juvenal*."* In 1794 she published her second novel, *The Widow, or a Picture of Modern Times*. In 1795 she published another novel, *Audley Fortescue*, under the name "Mr. Robinson."[16] In 1796 she published four major works: the novel *Angelina*; the novel *Hubert de Sevrac*; a sonnet series *Sappho and Phaon*; and the play *The Sicilian Lover*. The next year, 1797, saw two more novels: *Julie St. Lawrence* and *Walsingham*. 1798 was Robinson's first year since 1790 with no

[16] Or perhaps this is actually by a Mr. Robinson — I've seen it attributed both ways, and need to finish investigating.

publications, but in 1799, she perhaps made up for it by publishing two novels, *The False Friend* and *The Natural Daughter*, as well as a political treatise, *A Letter to the Women of England on the Cruelties of Mental Subordination.* She also began contributing weekly essays to the *Morning Post and Gazetteer* in 1799. After 1799 (and thus after this dissertation's decade of interest), Robinson's career and life is short. In 1800, she published her tenth novel, *Ellinda;* her verse collection *Lyrical Tales* in volume form as well as *The Mistletoe, a Christmas Tale* (as Laura Maria) in its own small book; and a translation from the German of Joseph Hager's travel narrative *A Picture of Palermo.* Robinson died in December 1800. After her death, her daughter Mary Elizabeth Robinson oversaw the posthumous publication of three more works: *Memoirs of the Late Mrs. Robinson, Written by Herself* (1801), a collection of poetry titled *The Wild Wreath* (1804), and a final collection of *The Poetical Works of the Late Mrs. Robinson* (1806).

Mary Robinson's early biography shares striking resonances with that of Charlotte Smith. Robinson was also married at the age of sixteen to a man who squandered his income. Robinson spent more than a year in King's Bench debtors' prison with her husband and infant daughter, from 1774 to 1775, from which she published her first volume of poetry. Rather than immediately taking up the mantle of a professional author, however, in 1776 she became an actress. As an actress, she met increasing success for four seasons — so much so that she attracted the attention in 1779 of the 17-year-old Prince of Wales, who determined to make her his mistress. After a period of semi-public flirtation, Robinson agreed, and retired from the stage. "Although the affair lasted less than a year, 'the Perdita' was notorious from then on; her gowns, her carriages, and her alliances became a constant source of discussion and speculation in the newspapers" (Mark Ockerbloom). The Prince of Wales had given her a bond of £20,000 at the start of their relationship, but when their relationship soon ended, he refused to pay. "She

demanded £25,000 for the return of the prince's letters. She apparently settled for £5,000, paid by George III ‘to get my son out of this shameful scrape.’ It was enough to stave off her creditors. In 1782, Mary obtained a further £500 annuity for herself, and a £200 annuity during the life of Maria Elizabeth, in return for the surrender of the Prince's bond.” (Mark Ockerbloom). Robinson continued to live in high society and the public eye, involved with Lord Malden, and then with Colonel Banastre Tarleton in a relationship which would be her longest. In 1783, Tarleton’s debts grew pressing, and his family attempted to sever his relationship with Robinson by promising to pay them if he moved to France without her. He left England in July 1783. Robinson, pregnant, borrowed money for his debts and set out alone to intercept him in Dover. The rough travel led to a miscarriage, the mistreatment of which ultimately paralyzed Robinson’s legs and left her with acutely painful rheumatism the rest of her life. For the next five years, Tarleton and Robinson lived variously in France, Germany, and England, variously together and apart. In 1788, Tarleton’s fortunes and Robinson’s health had recovered enough for her to move permanently back to England, and the two established neighbouring households on Clarges Street in London. Over the next decade, they continued their relationship, while Tarleton’s political career and Robinson’s literary one both flourished. In 1797, Tarleton ended the relationship, and in 1798 married a young heiress, to Robinson’s dismay. Her health worsened, and she died 26 December 1800, aged 43.

Robinson’s public stature fell sharply after her death. Behrendt’s comparison of posthumous reviews of Robinson’s work with reviews of Charlotte Smith’s work highlights the fact that “unlike Smith she was widely regarded as the instigator rather than the victim of her misfortunes” (192), due to the ‘immorality’ of her choices; as a result, her claims to sympathy and sensibility are dismissed by early nineteenth century reviewers. Behrendt argues that her

"notorious public behaviour seems to have ensured that she would not get anything like a fair hearing as *an artist* among the conventional critics of the time. Her case is a painful reminder of the pervasive power of a self-appointed coterie of predominantly male critics who considered themselves custodians of national public morality" (192). Susan Civale's work on Robinson's immediate nineteenth century reception bears out this dismissal of Robinson "as *an artist*" (Behrendt 192). Civale highlights the disparity between Robinson's poetry and fiction and her *Memoirs*: even after Robinson's other works had fallen out of print, the *Memoirs* continued to be reprinted, and to spur new writing in the form of "reviews, essays, spin-off novels, illustrations, poems, mini-biographies, entries in multibiographies, and citations in the life writing of other key figures" (194). It was Robinson's biography, rather than her writing, which "continued to interest, perplex, and charm readers" (Civale 194).

When literary scholars began to be interested again in Robinson in the 1990s, their work continued to be shaped by an intense awareness of her personal life and her theatrical career, often examining her works for the performance of Robinson's identity communicated within them. [For example, Judith Pascoe's 1997 *Romantic Theatricality* - *"*According to Pascoe, Robinson enacted in her writing an ever-shifting public identity" (Cross 6).] [Also Susan Luther, Stuart Curran, Eleanor Ty, and Sharon Setzer]. The theatrical performance of celebrity remains, as Ashley Cross argues, "the dominant lens through which her writing and her career have been interpreted" (6). The other major lens is Robinson's interaction with other Romantic poets, "in particular her relations with Coleridge and the Della Cruscans, to a lesser extent Wordsworth and, more recently, Southey and Smith" (Cross 6). This examination of the poetic Robinson nonetheless often employs the same framings as work on the theatrical celebrity Robinson, in emphasizing the strategic development of personae for public attention. Cross argues that "[t]he

publication of the eight-volume complete *Works by Mary Robinson* by Pickering and Chatto (2009–2010), under the direction of William Brewer, marked Robinson's official reentrance into the literary canon" (12), [but is she really "in" the canon as *an artist*?]

Chapter three does for Robinson the same initial process that chapter two did for Smith: I assemble, from a range of scholarly resources, an authoritative list of Robinson's publications, and seek them in each of the six databases. My initial hypothesis is that, due to Robinson's different posthumous reception, proportionally fewer of her works will be included in smaller archives compared to Charlotte Smith. In addition to comparing Robinson's inclusion to Smith's, I will use Robinson to explore the structural makeup of my databases, examining their representation of "authorship" the way I used Smith's *Elegiac Sonnets* to examine their representation of "editions." Robinson frequently wrote under pseudonyms or left works unsigned: how do the databases I examine handle authorship attribution? Discussing the digital encoding of Robinson's authorship raises the subject of authorship more generally, as a flexible concept which can resist strict encoding.[17] In particular, it draws attention to the importance of unsigned works, as distinct from anonymous works, and "open secrets" of authorship. As with Smith, I close the chapter with a consideration of Robinson herself as an author.

4.3. Ann Radcliffe

Chapter four of this dissertation examines the works of Ann Radcliffe across several digital corpora. Ann Radcliffe is selected as a contrasting model of popularity, as a famous and popular author with a relatively short list of publications in only one genre. Radcliffe is a particularly useful case study author to contest the relevance of "editions" as a marker of popularity: many of

[17] For example, several attempts at statistical authorship attribution have, rather than clarifying a singular author for a piece, instead pointed to the essential role of collaboration. In particular, I will discuss the interesting recent failures to identify the unknown negative reviewer of Coleridge's *Christabel*.

the later editions of her works had very large print runs, since the works were sure to sell, such that a decline in reissues is in fact a sign of success rather than failure. The precise details of these editions and their implications occupy much of chapter 4, and provide important complications for the models of popularity and legacy developed with Charlotte Smith and Mary Robinson in chapters 2 and 3. Radcliffe's first publication is at the beginning of this dissertation's decade of interest: in 1789, Radcliffe published her first novel, *The Castles of Athlin and Dunbayne*, which reached five London editions by the time of her death. In 1790, she published *A Sicilian Romance*, which reached six London editions. In 1792, *The Romance of the Forest*, which reached eight London editions. This novel earned her more positive notice and began the momentum for her career: it is after the success of The Romance of the Forest that her first two novels reached their second editions. In 1794, Radcliffe published her fourth novel, *The Mysteries of Udolpho*, which had two London editions that year, and reached twelve London editions by her death. In 1795, Radcliffe published her only non-novel work, the travelogue *A Journey Made in the Summer of 1794*, which saw two editions in London that year and a third before her death. In 1797, she published the last new novel to be released during her lifetime, *The Italian*, which reached a second edition before her death. After this dissertation's decade of interest, Radcliffe's works continued to be reprinted domestically and internationally nearly every year, but she ceased publishing new works. In 1802, Radcliffe wrote, but did not publish, *Gaston de Blondville*, which would finally appear in print posthumously in 1826. In 1816, a "new" publication appeared: an unauthorized volume *Poems by Mrs Ann Radcliffe*, anthologizing poems previously appearing in her novels. Radcliffe died in February of 1823. This prompted reprints of all of her works by S. Fisher in 1823, and by Ballantyne in 1824. In 1826, her husband published several pieces posthumously: *Gaston de Blondeville*; a poetic piece

St. Alban's Abbey …To which is prefixed a memoir of the author; and, in *New Monthly Magazine*, her unfinished essay "On the Supernatural in Poetry."

Radcliffe's biography, too, is short. Radcliffe never cultivated a public literary persona, which itself led to the development of a mythos of her seclusion. Almost all biographical information comes from Thomas Noon Talfourd's 1826 "Memoir of Radcliffe," written in close consultation with Ann Radcliffe's husband William. Radcliffe kept extensive journals, which Talfourd occasionally extracts, but her manuscripts were all destroyed shortly after her death, with the exception of forty-two pages of her commonplace book from near the end of her life, and a two-sentence letter to a Miss Williamson (Rogers 2). Deborah Rogers's 1996 *Ann Radcliffe: A Bio-Bibliography*, the first to examine the commonplace book, remains the best account of Radcliffe's life and major works. Some simple facts are known. She was born in London on July 9 1763, the only child of Ann Oates Ward and William Ward, a haberdasher who later managed a china shop. Both of her parents were close to their slightly more illustrious relatives, and encouraged an old-fashioned sense of gentility in their daughter (Rogers 3). Radcliffe was married in 1786 at age 23, to William Radcliffe, "a hardworking Oxford law graduate who became part-editor and owner of *The English Chronicle*" (Facer). Unlike Smith's and Robinson's husbands, William Radcliffe appears to have used the household's funds fairly responsibly, though when Ann Radcliffe's mother died, one stipulation of her will was that Ann's inheritance not be used to pay any of William's debts. Certainly, William Radcliffe promoted the image of himself as a nurturing helpmeet to his wife: in the 'origin story' of Ann Radcliffe's writing which Talfourd writes based on William's descriptions, Talfourd credits William with encouraging his wife's shy talents. Facer says, "He often came home late and in order to occupy her time, Radcliffe began to write, reading aloud the lines she had written during the day on his

return" (Facer). Given how many critics emphasize Radcliffe's obvious responsiveness to reviews — each novel directly altering the aspects most criticized in the previous — her writing seems not to have been *entirely* an idle amusement to pass long hours. Her six major publications appeared rapidly between 1789 and 1797. In 1798, Radcliffe's father died, leaving some of his property to Radcliffe, some to her mother, and a small amount to William himself should he outlive his wife (Rogers 11). In 1800, Radcliffe's mother died as well, leaving her property to Radcliffe on the condition that nothing be left to William and none of the money be used to pay William's debts (Rogers 11-12). In 1802, Radcliffe wrote, but did not publish, *Gaston de Blondville,* her last major work of writing.[18] Radcliffe spent the second half of her life enjoying the domestic retirement praised within her novels, sometimes travelling within England with her husband. She died in 1823, age 58.

In contrast to the brevity of Radcliffe's bibliography and biography stands the mountain of secondary writing on her works. [NOTE: This section remains incomplete, with only a rough sketch of the literature review in place.] During her lifetime, Radcliffe was widely reviewed and discussed. Scholars have often examined Radcliffe's responsiveness to reviews of her works, with each new novel changing the aspects which was critiqued most strongly in the previous.

---

[18] That Radcliffe lived another twenty-six years after the publication of *The Italian* with no further works is a fact which has apparently demanded explanation since the eighteenth century. "In the total absence of documentation, contemporaries were willing to believe, presumably because she was the reserved (female) author of Gothics, that Radcliffe was insane. … So reticent and self-effacing was Radcliffe that she never corrected rumors of her death or madness" (Rogers 13). Her husband, however, *was* interested in countering these rumours, and after her death in 1823 publicized a report from her doctor unequivocally stating that she had died of asthma and that her "mind was perfect in its reasoning powers (E131 104)" (Rogers 20)[cite the original], suggesting instead that Radcliffe had withdrawn from publication after her inheritance made the increased income no longer important to maintain the luxuries of her life. Scholars largely accept this financial explanation, and add to it the explanation that Radcliffe had always been a fundamentally shy person who found public criticism of her writing distressing. Although her works were widely praised, they received so much attention that inevitably there were also critiques, and she may have wanted to remain distant from "the parodies as well as the many inferior imitations of her work" (Rogers 13).

[So she ultimately produces Gothics that reviewers are willing to accept on their own moral terms — and therefore Gothics which subsequent generation are willing to accept/praise?]

At her death in 1823, despite the twenty-six years since her last publication, an outpouring of obituaries marked her as an esteemed literary genius. Much of this praise followed the template so famously set out in Walter Scott's 'Prefatory Memoir to Mrs Ann Radcliffe' in the tenth volume of James Ballantyne's Novelist's Library series. Despite Scott's dismissive attitude toward the other romances that beguiled his youth, "he nonetheless takes care to defend the unquestionable genius of Ann Radcliffe on at least three accounts: her presiding over 'a separate and distinct species of writing' (Scott 1824: xx); her ability to sustain her readers' interest and attention across three major novels … and her exploration of extreme human passions in appropriately southern European settings" (Townshend and Wright 5-6).

Since Scott, Radcliffe has continued to be consistently praised and studied, often in ways that establish the worthlessness of works similar to her; a 'token' who justifies the exclusion of women or feminine writing. In 2014, the volume *Ann Radcliffe, Romanticism and the Gothic,* edited by Dale Townshend and Angela Wright, commemorated 250 years since Radcliffe's birth. In this volume, Radcliffe has weathered multiple changes in the critical landscape" "Once-dominant psychoanalytic and feminist interpretations have given way to critical work drawing on political and religious history, print culture studies, theories of media and remediation, and new formalism" (Carson 127).

Chapter four uses Ann Radcliffe as an important contrast to Charlotte Smith and especially to Mary Robinson, to challenge the ease with which the number of unique titles can stand in for both "popularity" and "importance." Taking as my premise that Radcliffe was unquestionably both popular and important in the 1790s, I examine the extent to which she does (or does not)

rise to attention within my corpora of that decade. I explore both print runs and reviews as important supplementary forms of information, and prototype and test computational methods to integrate these kinds of information with more accessible metadata. Print runs are of particular importance to interpreting Radcliffe's career, since a large run of one edition can represent far more actual books produced than several small editions, and Radcliffe's celebrity often meant large print runs. To compare her output to Smith's and Robinson's, I test and compare a few methods of calculating the total number of books printed for each author during the decade. To compare her reception to theirs, I hope to use a prototype version of Megan Peiser's in-progress Novels Reviewed Database. In both cases, my method will begin with manually researching the three authors' major works, and then exploring scalable computational methods. The chapter closes, again, with a reflection on Radcliffe as an author.

4.4. Hannah More

Finally, chapter five of this dissertation examines the works of Hannah More across several digital corpora. Hannah More is selected as a particularly difficult writer to grapple with, both bibliographically and in terms of critical reception.

Hannah More's first book published was the pastoral drama *A Search after Happiness*, published in 1773. Her early writing is focused on the theatre. In 1774 she published her play *The Inflexible Captive,* which was produced in 1775 at the Theatre Royal in Bath. In 1776, she published *Sir Eldred of the Bower and the Bleeding Rock*. In 1777, she published her first conduct-book, *Essays on Various Subjects,* dedicated to the bluestocking Elizabeth Montagu. Her play *Percy, A Tragedy* was also produced at Covent Garden in 1777, running through 1778. In 1779, her third play, *The Fatal Falsehood,* was performed but was considered a failure, after

which More stopped writing for the stage. From this point, she primarily wrote various forms of didactic and religious prose, with the occasional didactic and religious poem. In 1782, she published *Sacred Dramas* and *Sensibility: A Poem*, her first foray into poetry. In 1783, she wrote *The Bas Bleu*, which circulated in manuscript but did not see print until 1786. In 1786 she published *Florio*, a poem in praise of rural life, and *The Bas Bleu*. In 1788 she published *Thoughts on the Importance of the Manners of the Great to General Society*, as well as *Slavery: A Poem*. More's first publication in this dissertation's decade of interest is her 1789 poem, *Bishop Bonner's Ghost*. In 1790, she published *An Estimate of the Religion of the Fashionable Works*. In 1793, she published "Village Politics," a counter-revolutionary tract for the poor which came to embody the reaction against Thomas Paine. In 1793, she published *Remarks on the Speech of M. Dupont* in aid of French emigrant clergy. In 1795 she began the massive undertaking of editing, and largely writing, *The Cheap Repository Tracts*, the writing for which she is now most known. These tracts were distributed as inexpensive chapbooks from 1795 to 1797. In 1798, she reprinted many of these tracts in volume form. In 1799 she published, under her own name, *Strictures on the Modern System of Female Education*, and another volume of Cheap Repository tracts. Unlike Charlotte Smith and Mary Robinson, More had a long career after this dissertation's decade of interest, nearly all in the form of didactic prose. A third volume of Cheap Repository tracts was published in 1800. In 1803, she published some patriotic ballads. 1804 saw "The White Slave Trade," "an attack on the frivolity of the fashionable world" (Stott). In 1805, she published, anonymously, *Hints toward forming the Character of a Young Princess*. 1806 saw another edition of tracts in volume form. In 1808 she published her only novel, *Coelebs in Search of a Wife*. In 1811 she published *Practical Piety*. In 1812 she published *Christian Morals*. In 1815 she published *Essay on the Character and Writings of St. Paul*. In

1817, in response to civil unrest, she began republishing the *Cheap Repository Tracts* as *Cheap Repository Tracts Suited to the Present Times*. She reprinted and distributed these from 1817 to 1819, and wrote some new ones. In 1819, she published *Moral Sketches* and an abolitionist poem. In 1821, she published *Bible Rhymes*. In 1825, she published *The Spirit of Prayer*, her last new publication, written at age 80 after more than 50 years of writing. A fifth volume of Cheap Repository tracts appeared in 1827. More died in 1833, age 88.

More's biography differs strongly from Smith's and Robinson's in that she never married. Instead, much of More's personal life centred around educational ventures. In 1758, when More was thirteen, her eldest sister Mary More opened a girls' boarding school, which Hannah More would soon assist in running. In 1762, this school moved to larger premises. In 1767, More accepted a proposal of marriage from William Turner, though this did not end in marriage: "After a rather humiliating six-year courtship (during which the pending marriage was postponed three times by the gentleman), a settlement was reached on behalf of More by her family, resulting in an annuity that granted her enough financial independence to embark on a literary career" (Prior). When the engagement was finally broken in 1773, More made her first visit to London, and her life's focus shifted from education to literature. She moved in literary circles, joined the Bluestockings, and particularly befriended the famous actor David Garrick. However, after Garrick's death and the failure of her third play, both in 1779, More's focus shifted again, from London literary life to religious life. Her interest in the growing movement of Evangelicals ultimately brought her to the Clapham Sect, whose foremost cause was the abolition of the slave trade (Prior). She returned to Somerset, buying her own house, Cowslip Green, in 1785, and religion brought her back to education. She and her sister Patty More founded the first of the Mendip Schools at Cheddar, Somerset in 1789. "With the encouragement and support of the

Clapham Evangelicals … More undertook opening and operating Sunday Schools … in order to teach the children of the laboring classes to read and write, to learn the catechism, and to lead moral lives informed by Christian teaching. Though the effort was clouded by political and theological controversy from various sides, More eventually opened sixteen schools that taught hundreds of students." (Prior). In 1790, the More sisters hand over the Park Street school to Selina Mills. In 1792, she founds women's benefit clubs at Cheddar and Shipham. 1799 marks the beginning of the "Blagdon controversy": "one of More's teachers is accused of Methodism; the accusation widens into a series of attacks on More for alleged religious and political subversion; as a result she suffers depression and nervous collapse and for a while is unable to write," until 1802 (Stott). In 1801, More retired— somewhat. She moved to a new home, Barley Wood, outside the village of Wrington, where she maintained a voluminous correspondence and continued to receive and advise many visitors (Prior). In 1813 she founded an auxiliary Bible Society at Wrington. "She continued to support generously efforts at poverty-relief until her death. More died September 7, 1833, having survived all of her sisters and most of the members of her London circle of friends" (Prior).

As with Robinson and Smith, More has presented scholars a challenge in the quantity and range of writing produced during her lifetime. "With a lifetime that spanned the Augustan age to the Victorian age, and a writing career nearly as long, Hannah More can be considered at once both the intellectual child of the Age of Johnson and, in spirit if not quite in chronology, the 'first Victorian'." (Prior) More's "pre-Victorian" evangelical leanings, and especially the Cheap Repository Tracts, have typically attracted the most attention. The Cheap Repository Tracts are a bibliographic challenge: it can be difficult to confirm the author, publication date, or circulation of any given tract, and there are at least a hundred tracts. Additionally, since the twentieth

century More has presented an additional challenge in her subject matter: as Scheuermann observes, "[t]he problem with More is that while she is a most interesting figure, her ideas are largely repugnant to modern sensibilities. Critics often deal with this inconvenience either by apologizing for their subject or by changing what she says so that she seems closer to us in spirit" (237).

More's literary afterlife began the year after her death, in 1834, with William Roberts' publication of a collection of her letters. Roberts' editorial choices set the tone for More's biographers by emphasizing her conservatism. His omissions may have improved More's standing in the nineteenth century, but by 1952 historian M. G. Jones already acknowledges "the unsympathetic portrait of her which has been handed down to posterity" (Jones 152) due to her approach to social reform. "Unable to refute these charges, Jones can offer only the partial excuse that More's attitudes were typical of her time" (Nardin 268). Unlike Smith and Robinson, More's reputation did not necessarily improve when she came to the attention of feminist scholars in the 1980s and 1990s, though work on More did increase. When Pederson writes about More in 1986, the previous scholarship that she cites is all historical: More matters to historians of 1790s politics, and to historians of the evangelical movement (Pederson 85), but Pederson sees little precedent for her own discussion of More's writing as having literary, rather than historical, importance. Pederson argues that "only by examining the Cheap Repository within the context of popular literature can we understand the tracts for what they were: a broad evangelical assault on late eighteenth-century popular culture" (88).

Pederson's attention to More's literary contexts is shared by other scholars of the 1980s, though only Mitzi Meyers, who in 1986 contextualizes More's didacticism alongside children's literature more generally, seems to form a positive opinion of More. "Mitzi Myers, delighted to

have found a female eighteenth-century writer who was clearly successful, largely rewrites More so that her 'didactic' works 'scarcely stand second to the canonical novel in interest and importance' in terms of 'what they reveal about women'" (Scheuermann 238). Elizabeth Kowaleski-Wallace's 1991 monograph, *Their Fathers' Daughters: Hannah More, Maria Edgeworth, and Patriarchal Complicity*, better sums up the consensus on More: complicit in patriarchy. Blanch's assessment is that "[d]espite the contemporary criticism that More was usurping roles that had traditionally been reserved for men, feminist critics such as Kowaleski-Wallace, Ellen Jordan, Mona Scheurmann, and Eleanor Ty have maligned More as antithetical to the course of social reform and have dismissed her as a passive agent of the patriarchy" (Blanch 87). More's next biographer, Patricia Demers, continues this assessment in her 1996 biography *The World of Hannah More*. "Patricia Demers argues in a recent study that More's "belief in a natural hierarchical social order," a belief which Demers finds "angering in its condescension and immobility," prevented More from doing anything significant to improve conditions among the poor (Demers, 2)" (Nardin 268). The emphasis on More's politics also begins to shift attention away from literary interpretation. "Writing for More is primarily a mode of instruction, whether in poetry, drama, essay, or tract, conveying either a female or a male voice of authority." (Demers 109)

In the early 2000s, scholarship emerges that is more ready to find merit or sympathetic politics in More. Scheuermann describes this phenomenon as critics "changing what she says so that she seems closer to us in spirit" (237); Scheuermann herself presents More as reprehensible propagandist. But it is work like Jane Nardin's — which argues that "although More was a less enthusiastic believer in the 'hierarchical social order' than most scholars have argued, the evasions and compromises she engaged in as a practical reformer helped to damage her

reputation with posterity" (269) — which takes the more common stance. This is the More who can be inspiring in her importance and innovation, as suggested by titles like Kevin Gilmartin's "Hannah More and the Invention of Conservative Culture in Britain" and Anne Stott's *Hannah More: The First Victorian*, both published in 2003.

Throughout this history of struggling with More's conservatism, More has been almost synonymous with the Cheap Repository Tracts. "At the time of the Repository's conception in 1794, More's public profile was that of a playwright and a controversial political commentator, yet, the annals of literary history emphasize her contributions as an educator and religious writer … Indeed, the 'sublime and immortal' Cheap Repository is the primary reason she is considered by some to have been the most prominent woman evangelical campaigner in England between 1780 and 1810" (Blanch 1). The history of these tracts begins with G.H. Spinney's 1939 bibliography of extant Cheap Repository tracts held by the British Library. Spinney's work filled an important need — "Scholarly recognition of the lack of clarity regarding the Cheap Repository bibliography, including authorship, dates back to Augustus De Morgan in 1864 (241-45)" (Blanch 11) — and has remained, eighty years after its publication, an invaluable source. It is perhaps only surpassed by Anna Maree Blanch's 2009 thesis, *A Reassessment of the Authorship of the Cheap Repository Tracts*. Blanch makes the case that "[s]tatements describing Hannah More's contributions have been repeated by generation after generation of scholars uncontested" (12), and describes with evident frustration several works which appear at first not to have based their claims on Spinney, but which, when traced, have their origin in the same single paragraph. Blanch herself has carried out new original research. Blanch identifies 127 tracts in the original run of the Cheap Repository Tracts, of which "58 tracts are either

conclusively or tentatively attributed to More and 17 to others, while 52 tracts are described as being yet unattributable to any particular author" (Blanch 85).

Chapter five takes the complex methodologies developed in the preceding three chapters, and stress-tests them on More's difficult publication history. The many Cheap Repository Tracts are difficult to comprehensively identify and attribute authorship to: how have digital resources dealt with these challenges? I am particularly interested in examining publication details which More might have exaggerated for promotional purposes: what grounds can be identified for her claims of exceptional circulation rates? Of particular interest is the fact that More reprinted some Cheap Repository Tracts in different formats for different price points and audiences. Can these differences be captured in the databases I examine? Examining More's oeuvre also raises questions of literariness: if, as I expect, the Cheap Repository Tracts are excluded from more selective databases, is this appropriate? What kinds of "non-literary" writing is included in each database? This question brings my bibliographic research back home to the core questions about canon-building which drive my inquiry. To this point, I have sought ways to evaluate authors' *popularity* and their *importance* in the print marketplace, as related metrics which underly the authors' reputations. With More, I explore directly the use of *literariness* as an inherently unquantifiable metric of selection.

5. Conclusion

In the following chapters, this dissertation weaves together an exploration of 1790s literature with an exploration of contemporary digital resources. As described in more detail above, chapter two uses Charlotte Smith to introduce the core differences between the databases under consideration, and to examine our conceptions of editions and reprinting. Chapter three uses Mary Robinson, whose printed output is expected to be largely similar in bibliographic terms to

Smith's (that is, they both consist of a large number of publications in a wide range of genres, which were successful enough to sustain a career as a writer), to examine questions about authorship and reception. Chapter four introduces an author who provides a contrast to both Smith and Robinson by producing a small number of titles in only one genre, Ann Radcliffe, prompting deeper inquiry into print runs and reviews, to develop richer metrics for "popularity" or "importance." Chapter five introduces another contrasting author, of the opposing type, Hannah More, whose large number of publications at varied price points presents a technical challenge for the tools I have developed, and also a theoretical challenge for the concepts of literature I explore. A brief conclusion synthesizes my findings on these four authors, and presents my assessment of how they relate to 1790s literature as a whole.

Chapter Two: Charlotte Smith and the English Short Title Catalogue

1. Introduction

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. In the coming chapters, I will draw materials from the English Short-Title Catalogue (ESTC), Eighteenth Century Collections Online (ECCO), the ECCO Text Creation Partnership corpus (ECCO-TCP), HathiTrust, and Project Gutenberg. As a concrete point of entry for each database, I explore which works by Charlotte Smith have been included in each database, and in what format they appear. I begin with an overview discussion of the databases which will form an ongoing object of study. I then describe how each database, as it grows more specialized, winnows down from Smith's full oevre to represent her works through different partial subsets. To explore the usable affordances of each database's holdings, I compare the accuracy of the text files produced by Optical Character Recognition (OCR) in both the ECCO and HathiTrust databases to the carefully hand-corrected transcripts contained in ECCO-TCP. Finally, I discuss in more detail the creation and structural assumptions of the ESTC, offering a case study close-reading of its implicit logic. These explorations lead to a broader analysis of how existing digital corpora do, and do not, serve to re-create the institutional function that John Guillory attributes to literary canons.

My examination of these databases will, of necessity, describe only 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 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 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 classifications[19] — infamously, many books have "1899" assigned as their publication date because this date was used as a placeholder for "no date". 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.

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.[20] Through its

---

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

[20] 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.

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, which provide a unique solution to real barriers for scholars of contemporary literature, 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,[21] 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 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

[21] I have heard it quipped more than once in conferences sessions that you always *think* that you're going to get your texts from OCR, but you always *do* get them from Project Gutenberg.

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.

2. Charlotte Smith in databases

For the purposes of this chapter, I briefly 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.

year	title	ed	ESTC	ECCO	Hathi	ECCO-TCP
1784	Elegiac Sonnets, vol 1	1st ed	ESTC yes	ECCO no	Hathi no	TCP no
1784	Elegiac Sonnets, vol 1	2nd ed	ESTC yes	ECCO yes	Hathi no	TCP no
1786	Elegiac Sonnets, vol 1	3rd ed	ESTC yes	ECCO yes	Hathi yes	TCP no
1787	Romance of Real Life	1st ed	ESTC yes	ECCO yes	Hathi yes	TCP no
1788	Emmeline	2nd ed	ESTC yes	ECCO yes	Hathi no	TCP no

1788	**Emmeline**	**1st ed**	ESTC yes	ECCO yes	Hathi yes	TCP no
1789	**Emmeline**	**2nd ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1789	**Ethelinde**	**1st ed**	ESTC yes	ECCO yes	Hathi yes	TCP no
1789	**Elegiac Sonnets, vol 1**	**5th ed**	ESTC yes	ECCO yes	Hathi yes	TCP no
1790	**Ethelinde**	**2nd ed**	ESTC yes	ECCO yes	Hathi yes	TCP no
1791	**Celestina**	**1st ed**	ESTC yes	ECCO yes	Hathi yes	TCP no
1791	**Celestina**	**2nd ed**	ESTC yes	ECCO yes	Hathi yes	TCP yes
1792	**Desmond**	**1st ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1792	**Desmond**	**2nd ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1792	**Elegiac Sonnets, vol 1**	**6th ed**	ESTC yes	ECCO yes	Hathi yes	TCP no
1793	**The Old Manor House**	**1st ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1793	**The Old Manor House**	**2nd ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1793	**The Emigrants**	**1st ed**	ESTC yes	ECCO yes	Hathi yes	TCP yes
1794	**The Banished Man**	**1st ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1794	**Wanderings of Warwick**	**1st ed**	ESTC yes	ECCO yes	Hathi yes	TCP no
1795	**The Banished Man**	**2nd ed**	ESTC yes	ECCO no	Hathi yes	TCP no
1795	**Rural Walks**	**1st ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1795	**Rural Walks**	**2nd ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1795	**Elegiac Sonnets, vol 1**	**7th ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1795	**Montalbert**	**1st ed**	ESTC yes	ECCO yes	Hathi yes	TCP no
1796	**A Narrative of the loss…**	**1st ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1796	**Rambles Farther**	**1st ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1796	**Marchmont**	**1st ed**	ESTC yes	ECCO yes	Hathi yes	TCP no
1797	**Elegiac Sonnets, vol 2**	**1st ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1797	**Elegiac Sonnets, vol 1**	**8th ed**	ESTC yes	ECCO yes	Hathi yes	TCP no
1798	**Minor Morals**	**1st ed**	ESTC yes	ECCO no	Hathi no	TCP no
1798	**Rural Walks**	**3rd ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1798	**The Young Philosopher**	**1st ed**	ESTC yes	ECCO yes	Hathi yes	TCP no
1799	**What Is She?**	**1st ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1799	**Minor Morals**	**2nd ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1799	**What Is She?**	**2nd ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1800	**Elegiac Sonnets, vol 1**	**9th ed**	ESTC yes	ECCO no	Hathi no	TCP no
1800	**Rambles Farther**	**2nd ed**	ESTC yes	ECCO no	Hathi no	TCP no
1800	**Elegiac Sonnets, vol 2**	**2nd ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1800	**What Is She?**	**3rd ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1800	**Rural Walks**	**4th ed**	ESTC yes	ECCO yes	Hathi no	TCP no
1800	**Letters of a Solitary Wanderer, vols 1-3**	**1st ed**	ESTC yes	ECCO yes	Hathi yes	TCP no

1802	**Letters of a Solitary Wanderer, vols 4-5**	**1st ed**	ESTC no	ECCO no	Hathi yes	TCP no
1804	**Conversations, Introducing Poetry**	**1st ed**	ESTC no	ECCO no	Hathi no	TCP no
1806	**History of England**	**1st ed**	ESTC no	ECCO no	Hathi no	TCP no
1807	**Beachy Head**	**1st ed**	ESTC no	ECCO no	Hathi no	TCP no
1807	**Natural History of Birds**	**1st ed**	ESTC no	ECCO no	Hathi no	TCP 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.

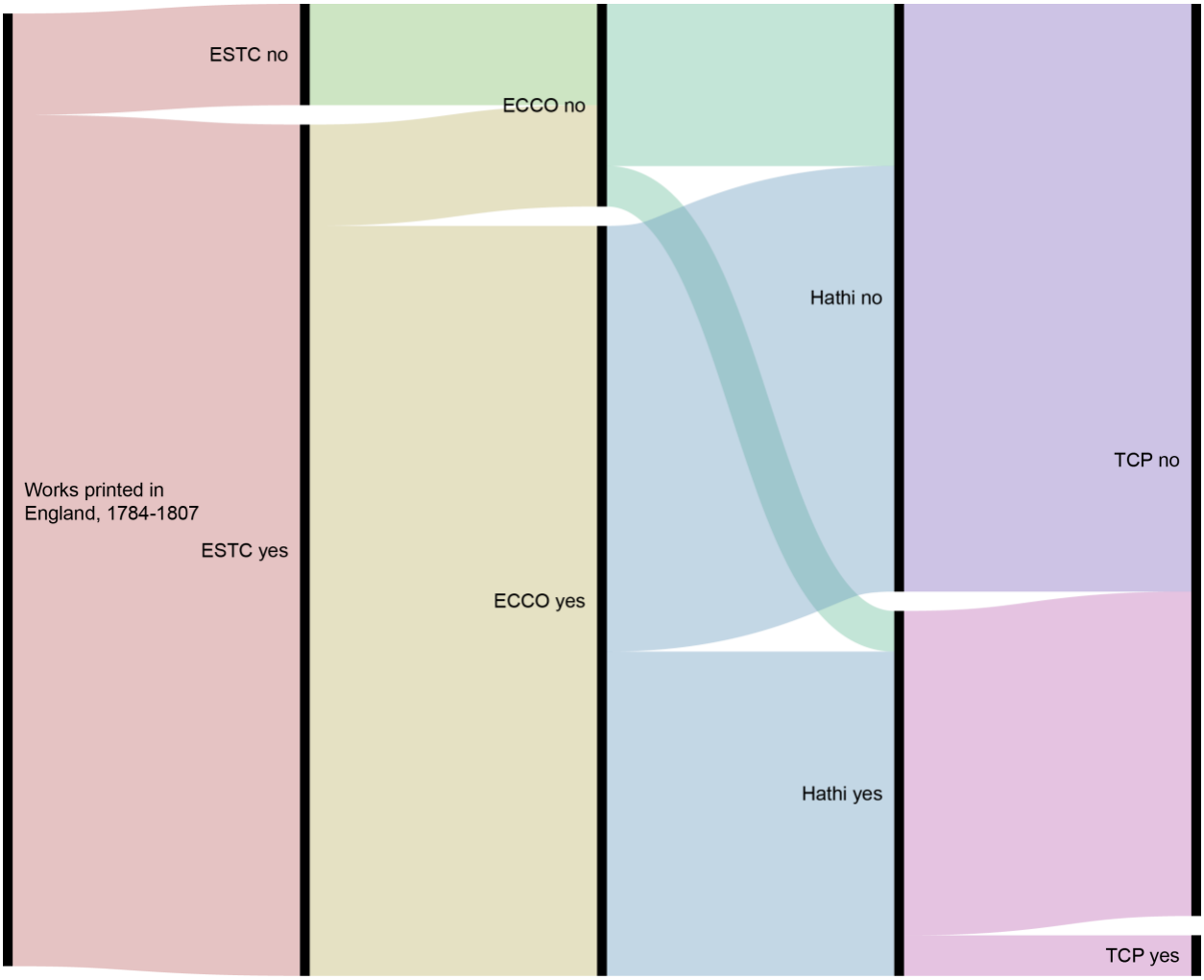


Figure 1: An alluvial chart, showing the winnowing down of Smith's works from database to database. Of the 47 editions printed in England between 1784 and 1807, 42 are included in the ESTC, and 5 do not appear in the ESTC because they were printed after 1800 and thus fall outside its purview. ECCO contains 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)[22]. 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.[23] The remaining HathiTrust included titles appear in both the ESTC and ECCO, and a further 21 titles appear as facsimiles in ECCO but not 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

[22] 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."
[23] 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).

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,[24] 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. ; 4º" with the 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.

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 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-based scholar, the path is easier for me to study Smith-the-children's-writer than other Smiths. It also demonstrates how questions of access, infrastructure, and selection are not just digital.

Only one of Charlotte Smith's works is available in Project Gutenberg: *Emmeline, the Orphan of the Castle* (first published 1788). This work appears in the ESTC and in ECCO in all its editions, but not in HathiTrust. It also, unfortunately, is not one of the two works selected for transcription by ECCO-TCP, so it is not possible to compare the accuracy of the Project Gutenberg text to a scholarly version. The Project Gutenberg copy does not state which edition (or editions) of the text were taken as its source, and does include several images which appear

---

[24] Later known as Primary Source Microfilm, an imprint of the Gale Group.

Victorian, suggesting that like many Project Gutenberg texts it is based on a late nineteenth century edition far removed from Smith's lifetime.

3. Reading OCR

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. A better test, of course, would compare a larger body of texts – though the obvious experimental design here would be to pair up texts from ECCO-TCP to their appearance in other resources, thus demonstrating the potential for ECCO-TCP's arbitrary selection of texts to perpetuate study of those texts in favor of others.

In comparing these three versions *of The Emigrants*, I proceed from the oversimplified assumption that the ECCO-TCP files are 100% accurate, and that any differences between the OCR and ECCO-TCP represents an OCR error.[25] 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 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, Juxta 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).[26] This surprised me, and suggests that skepticism of OCR in eighteenth century text mining may no longer be appropriate.

[25] One exception to this assumption has to do with treatment of the character ſ, which the TCP file modernizes to an s, but which HathiTrust renders as ſ. 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 ſ in HathiTrust to an s.

[26] Leaving the ſ characters unchanged in the HathiTrust document resulted in a .29 change from base (71% accuracy), so my normalization of ſ 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.

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 perfeetly 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 addreſſing a Com- poſition to you, which would never perhaps have exiſted, had I not, amid the beavy preſſure of many ſorrows, derived infinite conſolation from your Poetry, and ſome degree of animation and of confidence from your ºfteem. The following performance is far from aſpiring to be con- ſidered as an imitation of your inimitable Poem, “ The “TAsk;” I am perfºy fººl, that it belongs not to a feeble and feminine band to draw the Bow of Ulyſſes. The force, clearneſ, andſublimity of your admirable Poem; the felicity, almoſt peculiar to your genius, of giving to the moſt familiar obječís dignity and effečf, I could never hope to 3. – Reach

(HathiTrust)

TO WILLIAM COWPER, ESQ.
DEAR SIR,
THERE is, I hope, some propriety in my addressing a Com|position 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 con|sidered 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)

As this comparison shows, 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 ſ 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 a text file after the fact 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 ſ 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.

Figure 2 shows how Juxta highlights the words which vary between these three copies, and Figure 3 provides a visualization of where, across the text, there were more and fewer differences. This distribution of errors can be further illuminated by looking at the page images which were scanned to produce each OCR text: having seen, now, where the algorithm runs into difficulty, our attention is drawn to irregularities in the page and typography of the originals, reproduced in Figure 4 and Figure 5.

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 (above): 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.

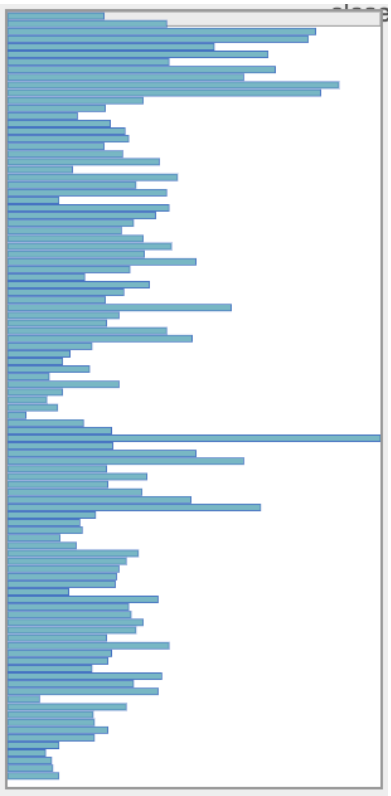
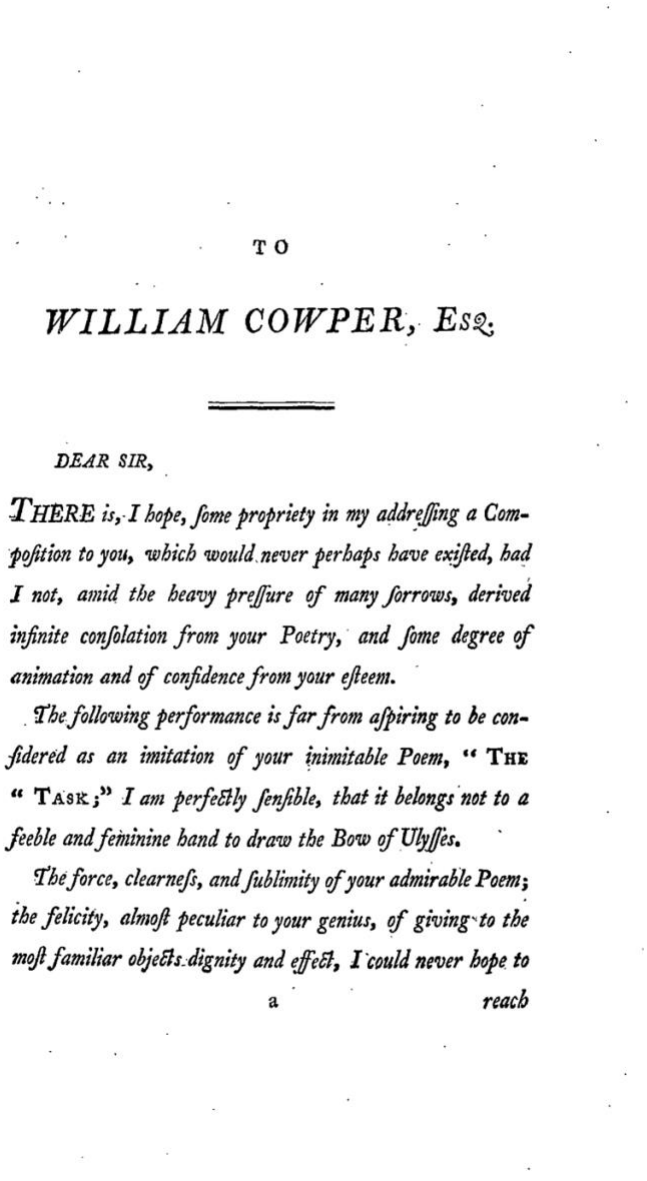


Figure 3 (left): A histogram, produced by Juxta, showing where the two ECCO and normalized HathiTrust witnesses show the most difference from the base ECCO-TCP copy. "Longer lines indicate areas of considerable difference, while shorter lines indicate greater similarity between documents." ("A User Guide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



TO

WILLIAM COWPER, Esq.

DEAR SIR,

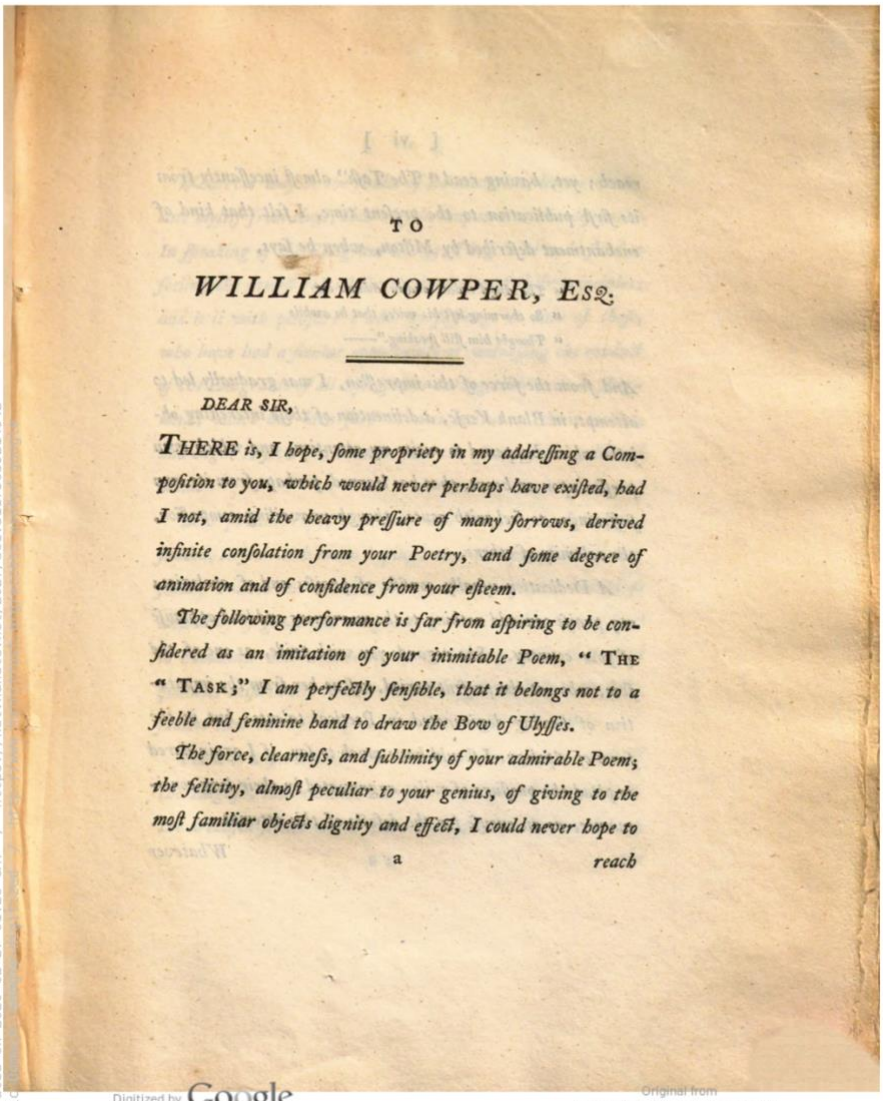
THERE is, I hope, ſome propriety in my addreſſing a Compoſition to you, which would never perhaps have exiſted, had I not, amid the heavy preſſure of many ſorrows, derived infinite conſolation from your Poetry, and ſome degree of animation and of confidence from your eſteem.

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

a *reach*

Figure 4: The facsimile of the first page of *The Emigrants* found in ECCO, which forms the basis of the ECCO OCR text.



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

a reach

Figure 5: The facsimile of the first page of *The Emigrants* found in HathiTrust, which forms the basis of the HathiTrust OCR text.

Examining the page images directly also reveals that the ECCO-TCP transcript has not been completely accurate, even though it has accurately captured each word and even preserved many aspects of the arrangement of the text on the page: the ECCO-TCP transcript silently amends the conventional inclusion of an additional quotation mark in the third line of the second paragraph, where the page reads “THE / “TASK”.

What this exploration of OCR demonstrates is the technical infrastructure which leads scholars to use one source rather than another, such that textual selection is actually superseded by *tool* selection. Even if a scholar might prefer to base their work on a fuller representation of Charlotte Smith's works than are included in ECCO-TCP, if the alternative is to have no usable source texts at all, each scholar is likely to make the pragmatic choice of not reinventing the wheel, and perpetuating the minimization of many facets of Smith's works. It also demonstrates, however, that despite a long conviction that OCR scans of eighteenth century texts are too degraded to be of use to any text mining research, the algorithm underlying HathiTrust is much more effective than the one which ECCO relies on, and if this technology were applied to ECCO's substantially more comprehensive and accurate collection, eighteenth century scholars could find ourselves with a sudden embarrassment of riches.

4. The history of the ESTC

The history of the English Short-Title Catalog is long, as befits its enormous scope. "The English Short-Title Catalog (ESTC) is a vast database designed to include a bibliographic record, with holdings, of every surviving copy of letterpress produced in Great Britain or any of its dependencies, in any language, worldwide, from 1473-1800" (CBSR). Today, "The English Short-Title Catalogue is the most comprehensive record of what has appeared in print in Britain and the English-speaking world for all branches of human experience from the last quarter of the fifteenth century to the start of the nineteenth. More specialized studies exist for fields and eras within that span, but no other resource matches ESTC's dependability over such a broad range" (Vander Meulen 265).

It began as the Eighteenth Century Short Title Catalogue in the 1970s, operating in a similar line as the original Pollard and Redgrave Short-Title Catalogue for 1476–1640, which first appeared in 1926, and Donald Wing's catalogue for 1641–1700, which appeared in 1951. These catalogues established the ambitious simplicity of the ESTC: to accurately describe every edition of every printed work in English or from the United Kingdom. After the completion of Wing's STC, "[e]xploratory studies, poorly funded and inadequate though they were" (Korshin 209) throughout the 1950s and 60s pursued the feasibility of systematically accounting for the much larger body of printed work produced in the eighteenth century. The Eighteenth Century Short Title Catalogue began properly in 1976, at a conference jointly sponsored by the British Library and the American Society for Eighteenth Century Studies (Crump 106). Here, "bibliographers and librarians attempted both to arrive at a consensus of the size of the task and the methodology that would have to be adopted to achieve a union catalogue. However, until the works were catalogued, it would not be possible to answer basic questions (such as the potential number of extant items) which would predetermine working methods. The very fact that they found it difficult to agree for want of sound and accepted figures indicated the need for ESTC." (Crump 105). A pilot project began at the British Library in 1977, under the direction of Robin Alston (Crump 105).

Unlike earlier Short-Title Catalogues, which appeared as lengthy print publications, the Eighteenth-Century Short Title Catalogue was conceived as digital from the beginning — a decision which, as Karian notes, "exhibited considerable foresight" (283) in the 1970s. As a result, "ESTC records existed in digital form long before many humanists saw computer technology as central to their work" (Karian 283). Robin Alston and Mervyn Jannetta developed their own cataloguing rules, distinct from the Library of Congress MARC and UK MARC

standards (Korshin 211). Once these standards were established, the British Library began to re-catalogue its own holdings, and in 1979 libraries in the United States, Germany, and Australia undertook to supplement them. In these international collaborations, "Where ESTC records already existed, these were adopted as the [new] record and only those works not held in the ESTC base file were catalogued again" (Crump 105). "One implication of the publication history of short-title catalogues is that they have been deemed functional and valuable even before they were complete. (That estimation is crucial, for their full completion is for all practical purposes impossible.) Judging that even a preliminary form of the records was useful to scholars, the planners of ESTC determined to conduct its development 'in full public view' and to make the incomplete file available 'warts and all' (in the words of Henry Snyder and Michael Crump, responding to criticism by Peter Blayney)" (Vander Meulen 270). Accordingly, the in-progress database "was soon available online, from 1980 via the British Library BLAISE [British Library Automated Information SErvice] system and from 1981 in the US Research Libraries Group RLIN [Research Libraries Information Network] system" (Norman). Each of these databases was worked on locally by researchers, and then updated and reconciled with each other weekly.

To supplement these databases, accessible almost exclusively to librarians with specialized training in operating them and primarily used by the scholars compiling the file, the ESTC intended to publish editions at particular milestones of completeness, intended for the use of non-librarian scholars. Their "first step, a fiche catalogue of [the British Library's] holdings, together with indexes, generated by the computer" (Crump 105) was published in a microform "snapshot" in 1983, but other milestones did not occur according to schedule. The "joint Anglo-American interim publication of the ESTC file " (Korshin 212) which was expected to follow on microform in 1984 (Korshin 212) did not appear. Alston attributed the delays partly to the

immensity of the task, and partly to the impact of short-term cost-cutting decisions, like the reduction of early-stage proofreading or of in-person examination of books, which dramatically increased the labour of verifying the resulting database record. Although he consistently warned "how easily strategic decisions based exclusively on cost usually lead to greater, not less, eventual costs" (Alston), the ESTC each year seemed to be facing a new budget struggle, and important maintenance labour was several times deferred. This created something like a paradox for the ESTC: funding bodies wanted to commit less money to a project which was behind schedule, but the project would remain behind schedule unless it was funded to complete the work required.

Nonetheless, work continued, and in 1985, the online databases in RLIN and BLAISE were upgraded to allow dynamic updates to a single shared file (Crump 106), which for the first time allowed continuous access to a shared record, rather than the constant exchange and messy merging of individual partially-overlapping records. "Until the file was dynamically available online on RLIN in 1985 batch processing was a weekly nightmare" (Alston). At this time, it was hoped that the new RLIN file would "result in a more complete and coherent 'first edition' of ESTC" to be published in 1989 (Crump 106), though this deadline, too, was not met. In the meantime "the ESTC file [was] available to scholars on both BLAISE-LINE and on RLIN." (Crump 106). To facilitate its use, the ESTC distributed "[a] simplified manual for searching the file on-line" (Crump 106). Crump took the opportunity of the update to rhapsodize on the database's potential usefulness for other scholars: "No longer is the scholar limited in access to the data by the fixity of the printed page" (106). This valuable resource was not without cost. Although the manual on how to formulate search queries was free, use of the ESTC itself was notably not. Institutions or individuals paid to subscribe to the ESTC itself, paid per query for

searches to be run, paid per minute for being connected to the database, and often paid for access to the computers they must use in their own libraries. Tabor says “the ongoing expense of consulting ESTC was the cyber-equivalent of the hefty up-front payment needed to acquire its printed predecessors, STC and Wing” (367).

“In 1987, with the agreement of the Bibliographical Society and the Modern Language Association of America, the International Committee approved the extension of the database to cover the period from the beginning of printing in the British Isles (ca. 1472) to 1700. The file changed its name to the 'English Short Title Catalogue', thereby keeping its well-known acronym. The USA team began cataloguing pre-1701 material in 1989, joined in the mid-1990s by the British Library team, and the resulting records were made available in the RLIN file from 1994.” (Norman). “In 1992, IESTC approved a further extension of the file to include serial publications. The USA team began work in 1994 on the cataloguing of serials within the scope of ESTC” (Norman).

Concurrently with the development of the ESTC, Wing’s seventeenth-century STC was undergoing redevelopment into a second edition, overseen by Katharine Pantzer. The second edition of Wing’s STC was published in two volumes in 1976 and 1986, followed by a set of exhaustive indexes in 1991. This second edition “represented a vast development of the original” (Vander Meulen 268), incorporating thousands of new entries, expanding the titles, and adding explanatory notes and headnotes. Its completion in 1991 also marked the end of the ability of its publisher and sponsor, the Bibliographical Society, to support it (Vander Meulen 269). “Accordingly, in 1999 the Society made an agreement with ESTC whereby the latter… would assume official responsibility for receiving new STC data” (Vander Meulen 270). The ESTC

continued to research new entries and improve existing ones, releasing a second edition of the file on CD-ROM in 1998 and a third edition in 2003 (Norman).

In 2006, almost thirty years after the commencement of the project, the ESTC underwent another major shift: the database was made publicly available to be searched for free online. This inspired more rhapsodizing, this time from Tabor: "The freeing of ESTC … now places in one location, for the consultation of anyone with internet access, the fullest and most up-to-date bibliographical account of 'English' printing" (367). At the same time, the ESTC began a project "to provide full title and imprint transcriptions for the eighteenth-century records" (Tabor 370). Vander Meulen says that "The history of ESTC is in fact the record of steady developments. Some have been conspicuous—for instance, the physical progression from a printed prototype to microfiche, CD, online access via the vendors Blaise Line and RLIN, and universal online availability through the British Library." (Vander Meulen 270) Many more have been less visible, in constant improvements to the accuracy and detail of the records. In 2011, the Center for Bibliographical Studies and Research at the University of California Riverside was awarded a planning grant from the Andrew W. Mellon Foundation to "redesign the ESTC as a 21st century research tool" ("Planning Grant"), which was followed in 2013 by a larger two-year grant to execute software improvements to the ESTC.

5. Reading the ESTC

One reason that it can be informative to close-read the data structures of a resource like the ESTC is that a resource's categories of knowledge are driven by the *uses* to which it expects that knowledge to be put. Examining the implicit assumptions that will make a given organization of knowledge seem logical, we can work backwards to the purpose of mission of the initial knowledge creation. Thus Tabor describes the data structure and the mission of the ESTC in a

single statement: “ESTC’s most basic bibliographical function is to provide, for each edition, a description of the ideal copy, meaning the most complete and correct manifestation of that edition as the printer and publisher intended it” (369). Korshin further elaborates the use envisioned for this information: “the ESTC's cataloguing rules have been devised in such a way that a scholar anywhere in the world can tell, from the ESTC entry, whether the copy of the book in his or her library is the same or different from the one listed in ESTC” (211). Both “edition” and “ideal copy” are terms defined around the interests of a specialist audience of bibliographers, which bear inexact but important relationships to the formulation of an ESTC record.

An “edition” is a group of copies of a work which are understood to be interchangeable with each other (Tabor 369),[27] though in practice different levels of granularity are applied in distinguishing between editions. The ESTC sometimes has separate entries for groups within an edition “when certain separately planned marketing units can be identified within the edition, such as reissues, imprint variants, and large versus regular-paper copies” (Tabor 369). Karian describes that “[s]ometimes the ESTC contains additional records if there are multiple *states* of an edition (a different state results from cancels or minor changes to the setting of type)” (289). Or, in “the later eighteenth century, when reprints from standing type became more common, ESTC cataloguers have occasionally granularized down to the level of individual impressions” (Tabor 369). As a result, Karian argues persuasively that ESTC records should not be treated as synonymous with “editions,” “issues,” or “titles,” since the same definitions of those boundaries may not be applied consistently. The specific question he poses is “What is the unit that the ESTC uses?” (289), and important question, to which the answer cannot really be “editions,” despite the best attempts of the ESTC bibliographers. Instead, he says “one should refer instead

only to the ESTC record, a unit created by the ESTC and having no meaning outside the ESTC" (Karian 289).[28]

The "ideal copy," too, represents an interpretation. Because the ESTC is essentially a model based on limited samples of an imagined lost prior whole — "the most complete and correct manifestation of that edition as the printer and publisher intended it," as Tabor termed it (369) — a new sample can change the model. As Tabor describes, "[a]s additional reports of copies arrive, it may be that the ideal description must change in response. For instance, the existence of a half-title may only emerge on the evidence of the seventh copy reported. A half-title would then be added to the description of the ideal copy, and the six previously matched physical copies will receive notes recording that they are imperfect in this respect" (370). The ideal copy, like the database itself, thus represents a moving target.

So, how do these ideas of the edition and the ideal copy shape the data structures employed in the building of the ESTC? Consulting an individual ESTC record in the online database, as we can see in Figure 6, reveals a lot of information all pointing 'outside' of the ESTC itself. It begins with six details which will be present for every title: the "System Number" and "Citation Number" uniquely identifying the record; the author; the title; the publication information; and a physical description. It then displays any uncategorized "notes," which in the case of *The Emigrants* (1793) consist of two additions to the physical description. The entry then points 'outward' to two "Surrogates": the microfilm, and the electronic reproduction of the microfilm

[27] "Because ESTC is a bibliographical database rather than a catalogue, strictly speaking, its records describe groups of copies," such as editions, "rather than specific copies," such as the Exeter Book (Tabor 369).

which is collected in ECCO. A very brief description is made of the work's content — its subject is "English poetry — 18th century" and its genre/form is "Poems" — which is the only information provided about the *work* rather than the *book*. The remainder of the record is an extensive list of libraries which hold copies of the edition, divided into the three geographic regions of "British Isles," "North America," and "Other," followed by a direct link to the ECCO copy referenced above.

---

[28] "The first problem relates to the unit of classification. A clearly defined unit is necessary to ensure that a study of change over time is reliable and based on consistent terms. What is the unit that the ESTC uses? Scholars sometimes answer by using the terms "edition," "issue," or "title" interchangeably. But since the ESTC does not rely in a consistent manner on any of these terms for its unit of classification, one should refer instead only to the ESTC record, a unit created by the ESTC and having no meaning outside the ESTC." (Karian 289)

Full Record

Permalink

Format options: Standard format Summary MARC tags HOLDINGS DETAILS

Record 1 out of 1 ← Previous record Next record →

ESTC System No.	006232050
ESTC Citation No.	T32633
Author - personal	Smith, Charlotte Turner, 1749-1806.
Title	The emigrants, a poem, in two books. By Charlotte Smith.
Publisher/year	London : printed for T. Cadell, in the Strand, 1793.
Physical descr.	ix,[3],68[i.e. 60]p. ; 4º.
General note	With a half-title.
	Numbers 9-16 omitted in pagination; text is continuous.
Surrogates	Microfilm. Woodbridge, Conn. : Primary Source Microfilm, an imprint of Gale Group, 2002. 1 reel ; 35 mm. Unit 355. (The Eighteenth Century ; reel 12409; no. 17). s2002 ctu b
	Electronic reproduction. Farmington Hills, Mich. : Thomson Gale, 2003. (Eighteenth century collections online). Available via the World Wide Web. Access limited by licensing agreements.
Subject	English poetry -- 18th century.
Genre/form	Poems.
Copies - Brit.Isles	Birmingham University Library
	Brighton Central Library
	British Library
	British Library
	Cambridge University Library (includes Sir Geoffrey Keynes Collection, British & Foreign Bible Society, & Peterborough Cathedral)
	National Library of Scotland
	Oxford University All Souls College Codrington Library
	Oxford University, Bodleian Library
	University of Warwick Library
Copies - N.America	Bryn Mawr College, Canaday Library
	Cornell University
	Duke University
	Harvard University, Houghton Library
	Henry E. Huntington Library and Art Gallery
	Library Company of Philadelphia
	McMaster University
	Miami University
	New York University
	Newberry
	Princeton University
	Temple University Samuel Paley Library
	University of Colorado
	University of Minnesota
	University of Texas at Austin, Harry Ransom Center
	University of Virginia
	Yale University, Sterling Memorial
Copies - Other	Alexander Turnbull Library
	National Library of Australia
Electronic location	Eighteenth Century Collections Online. Gale ;

Figure 6: A screencap of the ESTC record for Charlotte Smith's *The Emigrants* (1793).

This, however, is only how the ESTC *displays* its contents. Clicking another tab makes visible the MARC tags in which the data itself is stored. The MARC tags encode information at a slightly more refined level of detail. For example, the publication location in the standard view is listed as "Publisher/year" and displayed as the string "London : printed for T. Cadell, in the Strand, 1793." A human can parse that string, but as the MARC version of the same information reveals, it is made up of three points of information that have been combined. The MARC data is listed as "260," which is the MARC standard code for "Publication, Distribution, etc." The line itself is displayed as "|a London : |b printed for T. Cadell, in the Strand, |c 1793" — indicating three separate pieces of information in the subfields "a - Place of publication, distribution, etc.", "b - Name of publisher, distributor, etc.", and "c - Date of publication, distribution, etc." The separation of these points of information in the underlying MARC data is what allows the online database to conduct searches based on publisher, publication location, and date of publication. Even this is a reformatting of the underlying MARC code, which would read "##$a London :$b printed for T. Cadell, in the Strand,$c 1793" — with the two "#" symbols at the beginning encoding that this is the first edition.[29] It is, of course, only sensible for the ESTC to reformat its MARC code for display: MARC stands for MAchine Readable Catalogue, and machines and humans have very different needs as readers.

There are several different ways to search ESTC records. The "Search" button takes a user to the "Basic Search" function, from which there are also links to "Advanced Search," "Browse," and "Browse Libraries List" (which takes the user to the identical page as "Browse" but with "Library name" pre-selected as the index to browse). Once you have found a work of interest,

however, several new forms of searching become available, implied in the hyperlink formatting: almost any field in the entry can be clicked to reach other matching ESTC entries.

6. Conclusions

Like literary canons, these corpora — especially smaller ones, like the Eighteenth Century Collections Online Text Creation Partnership — are vulnerable to a critique of their selection methods on the grounds of representation. However, unlike the various changing literary canons of the past, digital corpora tend to conceal which particular titles have been selected as representative. I argue that Charlotte Smith's inclusion in these resources lags behind a scholarly consensus which sees her as increasingly important and canonical in the period. Her partial inclusion in ECCO-TCP seems particularly likely to lead to ill-supported conclusions by researchers who might easily assume that their text-mining research is taking her works into consideration. However, since none of her sonnets are included, nor any of the politically radical novels which made up a substantial portion of her latter career, *nor* any of her natural history, some of her most important contributions to the literature of the period are not able to impact studies in which they would be relevant. In particular, a study of women's writing through the lens of the ECCO-TCP would emphasize the most conventional and expected women's writing from Smith, with four volumes of one of her more straightforward marriage plot novels.

Exploring the technical affordances of the copies of Smith's works available in each database also shows why the distorted impression of Smith's works reflected in the ECCO-TCP's corpus is likely to persist and continue to be reproduced: without the foundation of a reliable but transformable text (in the form of a human-corrected transcription, rather than a page image or

---

[29] Technically, in the "##" sequence, the first "#" encodes that the work is a first edition (as opposed to a "2" for an "intervening" edition or a "3" for the "current" most recent edition), and the second "#" doesn't encode anything. That position in the MARC record is undefined, with no possible meanings, and simply

machine OCR), there is a nearly insurmountable technical barrier before any individual project. Even to assess the accuracy of the OCR texts in ECCO and HathiTrust, I must rely on ECCO-TCP. Guillory has already argued persuasively that representation in literary canons is a matter of selection, not of exclusion, that the default state for a given text is *not* to be included. For Guillory, this serves as a proof that sexism and racism are rarely the direct cause of a particular text lacking canonical status; the role of social oppression in limiting textual representation occurs before scholars make their choices, when classes of people are systematically excluded from the means of textual production in the first place, limiting what we may select from. In the case of digital corpora, also, I see that the rhetoric of "exclusion" is not accurate, and directs attention away from the more complex systems at play. Although I critique the failure of ECCO-TCP to include important and relevant works by Charlotte Smith, it does not seem that she has been *excluded* out of a prejudice against women's writing. Most likely, *The Emigrants* and *Celestina* were chosen because copies were conveniently accessible to a particular scholar involved in the creation of the ECCO-TCP, perhaps even directly related to a research question which would motivate them through the mind-numbing process of retyping long volumes of prose. Once these works had entered ECCO-TCP, they will naturally be re-used for text mining research which implicitly trusts the original selection. In this way, representation in digital corpora is a matter of infrastructure.

---

always contains a 'blank' #.

Appendix A: Codebase

1. gender-guesser.r

```
library(gender) # re-run these every time to get everything necessary installed
install_genderdata_package()
devtools::install_github("ropensci/genderdata")

testnames = c("john", "madison")

ECCOtestnames = c("Charles", "Society", "William", "René-Louis", "James", "John", "Edward", "Philip", "Hainault", "Junius", "Septimus", "Herodian", "John", "Benjamin", "Samuel", "Whitwell", "William", "Augustus", "Francis", "William", "Olaudah", "Candid", "John", "Charles", "William", "John", "Great", "Andrew", "Charles", "John", "Society", "John", "Western", "Elhanan", "John", "William", "Joshua", "Ann", "Charlotte")

gender(ECCOtestnames, method = "napp", years = 1789)
# if this makes an empty tibble, the code is probably fine, but I'm asking for something it has no data for (probably because the date is too early)
# ipums can go as early as 1789
# napp can go as early as 1758, but doesn't know the name "John" until 1769
```

Appendix B: Detailed Methods

1. ESTC sampling

My sliver of the ESTC was generously provided to me by the British Library in January 2017. It contains all items matching the query I specified, "(Words= alldocuments and W-year= 1789->1799 and W-Country of publica= enk)," which requests all documents published between 1789 and 1799 (inclusive) with a place of publication encoded as "England." Running this search on the ESTC website at the time returned 52,001 records. The tools used to create the file, according to the librarian with whom I corresponded, returned 51,965 records, 36 records having gone missing; however, the file itself contains only 51,860, another 105 mysteriously lost. These 141 missing records are currently an unsolved mystery. My records come from the British Library's ESTC database, rather than the STAR file. The corpus itself consists of a csv file[30] with fifteen columns of information. The columns are: "Type of resource" ("Monograph" or "Serial"); "ESTC citation number"; "Name" (e.g., of an author, editor or illustrator); "Dates associated with name" (generally, the years they lived); "Type of name" ("meeting/conference," "organization," or "person"); "Role" (e.g., "author," "cartographer," or "bookseller"), "All names", "Title", "Variant titles", "Place of publication", "Publisher", "Date of publication" (a single year), "Date of publication (not standardised)" (e.g., a year in roman numerals, or a date which includes a month or day), and "Publication date range" (for serials). In other words, it includes the very basic information of author, title, publisher, and year, in a complex structure which belies the apparent simplicity of these "basics." Some of the ESTC records included in this corpus do not necessarily match my selection criteria (England, 1789-99), which is inevitably true of every corpus collected, and which I discuss in more detail in SECTION, Data

Cleaning.

2. ECCO sampling

My first source of ECCO metadata consisted of MARC records, kindly provided by University of Toronto libraries (my thanks to Leslie Barnes!). I requested information for all works published 1789-99 in the UK (so, including Ireland and Scotland, but excluding America.)

My ECCO metadata presented particular challenges. I had access to MARC records, which stands for MAchine Readable Catalogue. At several points, I read this data with my feeble non-machine eyes in order to guide my data processing. Using MarcEdit, I converted these MARC records to csv files which could, in OpenRefine, be read, manipulated, and merged like my other corpora. Since I was not able to simply convert "all the MARC headings that exist" using MarcEdit, I used all numbers 1 to 999 and [will] delete empty columns. This process revealed that ECCO encodes much of its data in "unassigned" columns, rather than the standardized LOC categories.

I also have ECCO metadata records, now, from the Gale Digital Scholars Lab. Due to their restrictions on how much data may be downloaded at once, I have created 11 files, each one holding the metadata for all of ECCO's works in each year. ECCO-1789.csv, for example, is based on the following search: "LIMITS: Archive (Eighteenth Century Collections Online) And Publication Date (1789 - 1789)." These files initially contain works from a wide range of publication locations, not just England, since the Digital Scholars Lab does not provide any way to filter items by publication location.

3. OCR comparison

I started by testing just *The Emigrants*. To get the text without the XML markup, I viewed it

---

30 explain what a csv is

at https://quod.lib.umich.edu/e/ecco/004801766.0001.000?rgn=main;view=fulltext and simply copy-pasted the page into a plaintext file, manually deleting the header and footer website text so that the file only contained the poem. I saved it as emigrants-TCO-OCR.txt.

I attempted to download the HathiTrust edition by going to the "text-only view of this item" at https://babel.hathitrust.org/cgi/ssd?id=uc1.31175035214942;page=ssd;view=plaintext;seq=15;num=ix, where I discovered that I only had "one page at a time access to this item," even though it was correctly identified under "Rights" as a public domain work. Authenticating through the University of Toronto gave me access to the full work. This page came with the warning "Use of this online version is subject to all U.S. copyright laws. Please do not save or redistribute this file," which I disregarded under Fair Use to save a personal research copy of the text. Again the easiest method was to copy-paste the page into a plaintext file and delete unnecessary headers. I deleted "Book Text - Front Cover" and everything above, and "End of Section 9" and below, and saved the file as emigrants-Hathi-OCR.txt.

To download the ECCO OCR file, I went to the new Gale Digital Scholar Labs portal (since ECCO itself does not make the OCR available). The "basic search" very annoyingly attempted to autocomplete my search for "the emigrants" to unrelated terms, like "henry the eighth" and "female emigrants". The "advanced search" did the same, changing "the emigrants" to "therefore" and "mall of the emirates." Eventually, I was able to find the desired text by searching "the emigrants" as document title and "charlotte smith" as author. Gale Digital Scholar Labs prominently assigns an "OCR Confidence" of 95% to the record. The viewer did not give me 95% confidence. This method of accessing the text was the first to offer a download of the OCR. I downloaded it with their tool, examine the file, and deleted the disclaimer about OCR

which appeared at the top, then renamed the file emigrants-ECCO-OCR.txt to match the other files.

At this point I was ready to load the files into Juxta. I signed in to Juxta Commons (at http://juxtacommons.org/home/index) to use the web interface, and uploaded all three files as "sources." I prepared all three as "witnesses," then created a "comparison set" of "Emigrants OCRs" with them and collated. It took a long time to collate, and then I discovered that, unfortunately, whichever file appears first (in this case, the ECCO file) is taken as the "base" to compare the others to, and I couldn't determine how the base could be changed. I deleted the ECCO and Hathi witnesses from the set, then dragged them back in to it, to place them below the TCP witness. This caused unexpected server errors and failed, so I deleted the whole set, then created a new one with just the TCP witness and tried again. This worked, but the ECCO witness was immediately listed about the TCP witness, presumably due to alphabetization. I created a new witness from my TCP source, naming it beginning with A, and added that to the comparison set, then collated. 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 HathiTrust a .29 change from base (i.e., 71% accuracy).

Since some of HathiTrust's inaccuracy may come from the fact that it uses the ſ character whereas TCP modernizes this to an s, I then ran a second experiment. Using "find and replace" in TextWrangler, I changed all 1278 instances of ſ in the HathiTrust file to s, and saved this as a new file, emigrants-Hathi-OCR-regularized.txt. I uploaded this to Juxta as a new source and witness for the same set. In doing so, it became clear that Juxta takes whatever the newest witness is as the "base," so I had to delete/recreate my TCP witness again to make it the base. Once I got this working, the new HathiTrust copy had only a .09 difference from the TCP copy

— for 91% accuracy!

4. gender-guesser.r

The “gender” package in R is able to draw on a range of historical sources for gender information, but only one is applicable for this project. Most are US-based or only contain information beginning in the nineteenth century (or both).

If no value is specified, then for the "ssa" method it will use the period 1932 to 2012; acceptable years for the SSA method range from 1880 to 2012, but for years before 1930 the IPUMS method is probably more accurate. For the "ipums" method the default range is the period 1789 to 1930, which is also the range of acceptable years. For the "napp" method the default range is the period 1758 to 1910, which is also the range of acceptable years.

"The "napp" method uses census microdata from Canada, Great Britain, Denmark, Iceland, Norway, and Sweden from 1801 to 1910 created by the North Atlantic Population Project.” (Mullen, Blevins, and Schmidt) “For the "napp" method the default range is **the period 1758 to 1910**, which is also the range of acceptable years.”(Mullen, Blevins, and Schmidt). “The North Atlantic Population Project (NAPP) is a machine-readable database of the complete censuses of Canada (1881), Denmark (1787, 1801), **Great Britain (1851, 1861, Scotland 1871, 1881, 1891, 1901, 1911)**, Norway (1801, 1865, 1900, 1910), Sweden (1880, 1890, 1900, 1910), the United States (1850, 1880) and **Iceland (1703, 1729, 1801**, 1901, 1910).” (NAPP)

“The "ipums" method looks up names from the U.S. Census data in the Integrated Public Use Microdata Series.”(Mullen, Blevins, and Schmidt) “For the "ipums" method the default range is **the period 1789 to 1930**, which is also the range of acceptable years.”(Mullen, Blevins, and Schmidt)

To get a handle on how the gender package in R actually worked, and to assess how well it

would meet my needs for this project, I began by making a sample csv of 41 arbitrary titles from ECCO's 1789 holdings. The choice was driven by simplicity for a proof of concept: my ESTC and full-ECCO files were too large to open on my computer in spreadsheet software, and writing a program to extract a random sample was more work than it was worth, so I chose a file that I knew I could open in Numbers, the sample of all 1789 ECCO titles which I had recently downloaded through Gale's Digital Scholar Lab. I selected some rows from the top of the list, skimming the names as I went to make sure I had selected enough titles that they would include two by women, and then copied those rows to a new spreadsheet (ECCO-1789-sample.csv). After some false starts with RStudio, I was still struggling to read the file into the program. Since all I wanted to know was whether, once I got it running, the gender package would tell me results worth working with, I decided to do this first test in whatever way would get a result, even if it required a great deal of non-scalable manual labor.

I copied the authors column into a plaintext file, and manually created a comma-separated list of just the first name of each other, to match the data format which the gender package required. (I took the first word before a space in each name, to reflect how a future first-name-grabber algorithm would work, even when this meant choosing things which were clearly not first names — one of my core questions is how the gender package will handle those kinds of exceptions.) The resulting list of names was saved as ECCO-1789-sample-firstnames.csv. Trying to paste in this list for the names finally helped me understand what is meant by a "character vector" in R. It's basically an array of strings. You have to make the character vector before you can run gender() on it. Sample code:

```
names = c("john", "madison") # creates a character vector called "names" by 'combining (c-ing) two strings
gender(names, method = "demo", years = 1985) # guesses gender for both of those names
```

Once I had gotten something to produce guesses for more than one name, I wanted to try using a method other than "demo." This sent me down a spiral of trying to install genderdata, and trying to install devtools to install genderdata. Finally I got everything installed and able to run!

If a command is run requesting a name-year-method combination for which the package has no data, it simply returns an empty "tibble." The following three sets of results illustrate the limits and possibilities of napp and ipums data:

earliest year possible with ipums:

```
> gender(ECCOnames, method = "ipums", years = 1789)
# A tibble: 25 x 6
   name      proportion_male proportion_female gender year_min year_max
   <chr>               <dbl>             <dbl> <chr>     <dbl>    <dbl>
 1 Andrew                  1                 0 male       1789     1789
 2 Ann                     0                 1 female     1789     1789
 3 Benjamin                1                 0 male       1789     1789
 4 Charles                 1                 0 male       1789     1789
 5 Charles                 1                 0 male       1789     1789
 6 Charles                 1                 0 male       1789     1789
 7 Charlotte               0                 1 female     1789     1789
 8 Edward                  1                 0 male       1789     1789
 9 Francis             0.464             0.536 female     1789     1789
10 James                   1                 0 male       1789     1789
# … with 15 more rows
```

earliest year with napp that knows about "John":

```
> gender(ECCOnames, method = "napp", years = 1769)
# A tibble: 7 x 6
  name  proportion_male proportion_female gender year_min year_max
  <chr>           <dbl>             <dbl> <chr>     <dbl>    <dbl>
1 John                1                 0 male       1769     1769
2 John                1                 0 male       1769     1769
3 John                1                 0 male       1769     1769
4 John                1                 0 male       1769     1769
5 John                1                 0 male       1769     1769
6 John                1                 0 male       1769     1769
7 John                1                 0 male       1769     1769
```

more direct comparison between napp and ipums

```
> gender(ECCOnames, method = "napp", years = 1789)
# A tibble: 9 x 6
  name      proportion_male proportion_female gender year_min year_max
  <chr>               <dbl>             <dbl> <chr>     <dbl>    <dbl>
1 Charlotte               0                 1 female     1789     1789
2 John                    1                 0 male       1789     1789
3 John                    1                 0 male       1789     1789
4 John                    1                 0 male       1789     1789
5 John                    1                 0 male       1789     1789
6 John                    1                 0 male       1789     1789
7 John                    1                 0 male       1789     1789
8 John                    1                 0 male       1789     1789
9 Samuel                  1                 0 male       1789     1789
```

Neither is very stunningly thorough, especially since the year is supposed to reflect the *birth* year of the person, and I can only barely get information about the year they published something. Somewhat to my surprise, the ipums method — which I originally planned to reject in favor of napp, because ipums is US only — has much more data available, and doesn't seem to reflect any US-specific oddities. So, I might try to use both ipums and napp, but if that is difficult, just using ipums seems appropriate. I'm not entirely sure, yet, whether using ipums will be faster than just manually assigning genders: my next question, I think, is determining how many unique first names there are, and then trying to guess how many of them ipums would be able to sort out for me. It managed 25/41 = 61% of the random ECCO sample, which is a substantial number, especially since it will by definition include the most common names. Expanding the acceptable years slightly also seems to help it know a lot more about the names, while still according with my own estimates. (ipums can learn "Augustus" in 1799, for example.)

Appendix C: Tables and Figures

Table 1

Table 1 shows data that I compiled by hand in Numbers. The first three columns are based on my synthesis of scholarship on Charlotte Smith. As I consulted a range of work on Smith, I updated this information to reflect the most complete and accurate information possible. My editorial decisions included, for example, the exclusion of *D'Arcy* from consideration, since it was never published in England. I introduced standardized titles for the two volumes of *Elegiac Sonnets*, retroactively naming the initial publication "volume 1" to distinguish it from the second volume which would appear 13 years later, so that each title has its own edition count. The next four columns represent the results of my queries in the ESTC, ECCO, ECCO-TCP, and HathiTrust databases. I searched each database with several queries to locate Smith's works, beginning (where possible) by finding all works categorized under her authorship, and then searching individual titles of works. This figure shows the simplified results from a more detailed spreadsheet, which also includes links to the records themselves where they exist, and notes on how the records are encoded (e.g., multiple volumes or all as one volume.) Simplifying inclusion down to a boolean yes/no involved some editorial decisions. If only *part* of a work was included (as in HathiTrust's record for the first edition of *Celestina*, which only includes volumes 3 and 4), I recorded that as a "yes." If a work was only included in its Dublin edition (as in HathiTrust's record for *Desmond*), I recorded that as a "no." These searches were conducted in February 2020.

Figure 1

Figure 1 is based on the data recorded in Table 1, which was pasted into the RAW Graphs

visualization tool (Mauri et al.) and processed using their default settings for an alluvial diagram. Using an alluvial diagram required imagining the databases as sequential "stages" through which all books flow. Accordingly, I added a "source" for this flow of books by adding a column labeling every edition as falling into the category "Works printed in England, 1784-1807." I could have chosen to place the following "stages" in any order; to assist in visualizing Smith's representation in these databases as a process of winnowing down, I chose to place them in order from largest selection to smallest. The scale of each "strand" of the diagram is scaled in width based on the number of editions it represents, as per RAW Graphs' default settings. The colours, fonts, and width of the graph are also simple defaults.

Figures 2-6

The other five figures included are screen captures taken from my browser's display of the information described in each figure's caption, with no further intervention.

Works Consulted

Alang, Navneet. "Literature is Not Data! But Data is a Way to Read." *Hazlitt*, 7 Nov 2012. web.archive.org/web/20191023050702/https://hazlitt.net/blog/literature-not-data-data-way-read.

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.

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.

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.

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.

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.

Chun, Wendy Hui Kyong. “Queerying Homophily.” *Pattern Discrimination*, by Clemens Apprich, Wendy Hui Kyong Chun, Florian Cramer, and Hito Steyerl, meon press 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.

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.

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.

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.

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.

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.

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.

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.

HathiTrust. “Getting Content Into HathiTrust.” *HathiTrust*, web.archive.org/web/20190915204356/https://www.hathitrust.org/ingest. Accessed 15 Sept 2019.

—. “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.

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.

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.

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.

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. Uboldi, 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/.

McLaverty, 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.

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, issu 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.

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.

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.

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 Kosofky. "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.

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.

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.

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. T*he 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.

—. “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.

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.