Print Politics
in the Digital Archive,
1789-99

Lawrence Evalyn
August 4, 2020

Committee:
Alexandra Gillespie (Supervisor)
Terry Robinson (Co-Supervisor)
Tom Keymer

First two chapters of a dissertation in progress, submitted in support of an application for the Doctoral Completion Award.

Table of Contents

Introduction

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

[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."

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, 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, and HathiTrust. 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. For each database, I download their holdings identified as printed in England 1789-1799. I identify how many titles the database attributes to each year. I calculate how many works are attributed to male, female, or unknown authors.[2] These very simple pieces of information, when they differ widely between databases, 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.[3] 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 under-representing novels. Or it could include a representative number of novels, but disproportionately include novels by men. My investigation allows me to identify the patterns of selection. To ground my analysis in specifics, I take Charlotte Smith as a case study author. Smith has a long history of contentious reception, rooted in debates about seriousness, popularity, and women's writing. I revisit them to see how her career and reception might be interpreted through a new lens. I do so, in part, to

[2] 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.

[3] 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.

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

1.1. From Canon to Corpus

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 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.[4] 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

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

exclusivity" (133). Under this system, 'serious' literature may not be identifiable linguistically, but it can still be identifiable by the difficulty of accessing it. This history of canonization has important implications for the field of literary study. As Guillory himself insists, if the aesthetic value of a text is determined by the social operations of class, it undermines the notion of literature itself as a category of writing distinguishable in aesthetic terms from non-literary writing. Guillory's book is motivated by the canon debates of the 1990s, which were driven by an urgent re-valuation of literature by women and people of colour.[5] 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

[5] 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.

be the solution" ("Conjectures" 55) is the problem of conceiving of *world* literature, rather than the "canonical fraction, which is not even one per cent of published literature" (55). His new methods are meant to enable literary studies to examine a new object. The field of distant reading has been moving away from Moretti himself. However, it is still shaped by the attempt to redefine the disciplinary domain of literary studies. In many cases, the new domain is no longer the "canon" but the "corpus," a collection of texts which are studied *en masse* for macroanalytical insights. Katherine Bode, for example, in "The Equivalence of 'Close' and 'Distant' Reading," argues that Moretti and Matthew Jockers replicate the approaches of New Criticism with their corpora, and calls for "a new scholarly object of analysis" (79) that directly examines historical and textual context of corpora as representations of "literary systems" (97). Lauren Klein, too, treats the textual corpus as the new object of literary analysis requiring curation, contextualization, and interpretation. Her critique argues that "it's not a *coincidence* that distant reading does not deal well with gender, or with sexuality, or with race," but also that these failings are not inevitable: "it's not that distant reading *can't* do this work," she insists, "it's that it's yet to sufficiently do so" (n. pag.). Bode, too, despite her strong critique of distant reading as it has been practiced by Moretti and Jockers, does not blame distant reading itself. Distant readers like Moretti and Jockers, she argues, "while claiming direct and objective access to 'everything,' … represent and explore only a very limited proportion of the literary system, and do so in an abstract and ahistorical way" (78). Klein, like Bode, calls for "more corpora—more accessible corpora—that perform the work of recovery or resistance" to allow research "beyond quote 'representative' samples, which tend to reproduce the same inequities of representation that affect our cultural record as a whole" (n. pag.). This framing re-creates, at the site of the corpus, the identical narratives of exclusion and representation which were previously located in critiques of the canon.

The relocation of the debate from the canon to the corpus is not without grounds. As this dissertation will explore in depth, challenges to the technological accessibility of texts have created new hierarchies, and a new “great unread.” Each archive represents a unique set of choices in response to the same sets of questions: what to include, why, how; what to make accessible, why, how, to whom; what, in the end, makes a text matter, and what we are meant to *do* with texts. For example, the English Short Title Catalogue records 51,965 titles printed in England between 1789 and 1799. The corpus most commonly used for DH work on eighteenth-century literature, ECCO-TCP, includes only 466 titles for that same time period. What are the other 51,499 titles, why are they accessible in the ways they are, and what does it mean for digital eighteenth-century studies that they are not included? Although the examination of databases prompts similar hypotheses of exclusion as in longstanding conversations about canons, digital databases do not simply replicate new canons. [By the end of my dissertation, I will be able to state here what IS happening — something structured by related logics of access and prestige, and related simplifications of historical complexity, and related *institutional* replication of privileged texts.. But very importantly different, too, since we don’t *read* databases.] In a series of computational and non-computational research processes, I examine six databases of eighteenth-century texts to learn about four eighteenth-century authors, and I examine four eighteenth-century authors to learn about eighteenth-century databases. This dissertation, therefore, takes place within three scholarly conversations: the digital humanities, as an increasingly self-reflective set of practices; eighteenth-century studies, and the challenges presented by the 1790s; and the frameworks of reparative reading within queer theory which seem to offer valuable resources for both. The remainder of this chapter will describe in more detail the relevant scholarship shaping my frameworks, and then introduce my chapters by introducing my four case study authors.

1.2. Frameworks

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

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

The theoretical frameworks of this dissertation are drawn from the fields of feminist DH and queer DH, and from non-DH schools of thought which seem to offer valuable tools. My core motivating framework, as I conceptualize my work, is that of reparative reading. Eve Sedgwick's "Paranoid Reading and Reparative Reading" persuasively describes in the dominance of paranoia in literary criticism, and attempts to sketch an alternative in what she terms reparative reading. A paranoid rhetoric of exposure and critique strikes me as the most obvious narrative to structure this dissertation's investigation of the uneven institutional valuation of different writing. 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.[6] 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

[6] 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.

they are. Even deleting or ignoring information is still a way of "accounting for" it in the coding process: some part of the program will have to say, in effect, 'if I get an input that doesn't match what I expect, discard it.' Choosing to ignore contradictory or difficult information carries the assumption that this information does not 'count,' or does not matter to the question at hand. The choice faced by scholars is how to address our encoded assumptions. The encounter with variety does not in itself produce nuanced results: it is possible to selectively ignore any uncomfortable details. But it is also possible to do computation reflectively, asking not "how can I make this work the way I want?" but "where do my assumptions encounter resistance?" and turning one's attention to the nature of the resistance. Integrating this reflection into the research process can allow a scholar to avoid both the pitfalls of "conquering" their material and of claiming an algorithmic grasp of "objective" truth.

"The relationship between the individual cultural object and the curated dataset is not a transparent one; the latter is rather a heavily mediated and discipline-specific representation of the former. Through the collection and curation of our own dataset we are acutely aware of the choices that went into its creation. The use of already curated datasets has other undeniable advantages: it may temper the influence of the researcher on his or her findings; furthermore, from a practical standpoint, it allows work to advance past the time-consuming labor of curation. While we would not suggest that researchers need to reinvent the wheel, we do advocate for a more explicit reflection on the relationship between the dataset and the objects it describes. Such reflection allows for a deeper resonance between digitally enabled research agendas and existing intellectual and disciplinary traditions." (Vareschi and Burkert 612)

To bring these principles into the field of Digital Humanities by way of an example, I want to offer an alternative geneaology 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 Thomisticus* is often credited as the first large scale computational literary study.[7] Rather, they emphasize, Miles' origin story for computational literary study "can stand as an example of how we might write a history of literary scholarship that does not center originality and individual accomplishment" (n. pag.). Unlike Busa, Miles not only gave authorship to the (female) graduate students who carried out much of the labour of creating the concordances, she also thanked and credited the (female) punch card operators who encoded the resulting data.[8] Moreover, when talking of Penny Gee, one of the female staff members of the

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

[8] 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

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 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

Penny Gee.

be incapable of human fallibilities like racism. But as Safiya Noble shows extensively in *Algorithms of Oppression*, algorithms by default reproduce, and can easily exaggerate, the assumptions and biases of the culture in which they are made (CITE). In other words, in a racist world, algorithms *are* racist — and sexist, and duplicative of all other systemic inequities. To analyze an algorithm, one must articulate the implicit argument underlying the assumptions that allow it to operate — what Ian Bogost would call its procedural rhetoric. As Katherine Bode's recent work on data-rich literary history has shown, digital infrastructures themselves contain an implicit procedural rhetoric, even an argument, which must be addressed.

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

1.3. Methods

This dissertation undertakes computational distant reading. At every possible point, however, the underlying methodology will be made visible, and its assumptions scrutinized. The bibliographic histories of my multiple corpora are explicit objects of inquiry. Much of the code underlying this project I have written myself. Some has been written at my request. In every case where the code is available to me, the program itself appears in Appendix A ("Codebase"), accompanied by a plain language explanation of how it operates. Where I have used closed-source software, Appendix A contains an explanation of my best guess at its underlying process. My exact use of these tools — sufficient for another to replicate my work — is provided in Appendix B ("Methodology"). These details are explicated in full in the appendices in order not to over-burden the body of the dissertation, but they are by no means *confined* to the appendices. Computation is not a "black box" to be consulted for simple answers, but is inextricable from my reasoning and argument.

My attention to the *sources* of digital knowledge creation comes, in part, from Johanna Drucker, and her distinction between "data" and "capta." Drucker, in "Humanities Approaches to Graphical Display," specifically addresses the digital humanities practice of creating, and then close reading, data visualizations. She argues that the tools for visual representation which may be effective in the sciences cannot be simply and uncritically transposed to humanistic subject matter. When an experiment is presented as a 'data visualization,' she says, "the rendering of statistical information into graphical form gives it a simplicity and legibility that hides every aspect of the original interpretative framework" (8). In fields where the readers of such charts are also frequent creators of charts, and where norms exist to explicitly describe one's interpretive frameworks in a methodology section, the simplicity and legibility of an individual chart may be a benefit which does not impede complex scrutiny of the information it presents.[9] In a field like

[9] It may also be the case, of course, that even fields with a long history of graphical display would

literature, however, the “graphical force” of something like a network graph or even a simple pie chart “conceals what the statistician knows very well — that no ‘data’ preexist their parameterization” (8). Drucker problematizes the term “data,” the etymology of which presents it as a “given” which is stable and independent of observation. She proposes that humanities visualizations embrace, instead, the framework of “capta,” that which is “‘taken’ actively” (3), “fundamentally codependent, constituted relationally, between observer and observed phenomena” (50). Drucker’s assessment shapes my own prioritization of qualitative and 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.[10] 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

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.

[10] 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.

of the engineers' report at NASA which led to the ill-fated launch of the Challenger space shuttle in 1986: as his extensive visual analysis argues, the engineers (untrained in graphic design) unintentionally obfuscated crucial information about the day's launch conditions. The poorly designed graphics these engineers produced made the launch appear low risk to their superiors; despite the engineers' strong warnings, their verbal argument was disregarded in favor of their accidental graphical argument. As Tufte demonstrates, a few simple alterations of their graphic design would have made it obvious that the day's unprecedentedly low weather was extremely dangerous, and potentially averted disaster.[11] 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 \cite{Cairo:ikIksuMr}. His preceding book, *The Truthful Art*, addresses the creators of good faith infographics with insights into visual manipulation \cite{Cairo:2016uv}. Cairo draws a distinction between "data visualization" and "infographics": "an infographic tells the stories that its designer wants to explain, but a data visualization lets people build their own insights based on the evidence provided," summarized more succinctly as "infographics to explain, data visualizations to explore" \cite{Cairo:2014tl}. Using this terminology, my argument will proceed with infographics in the body of the dissertation as curated figures to support my argument, with fuller data visualizations available in Appendix C ("Data") to allow further exploration. Following in both Tufte and Cairo's footsteps, I conceive of the figures throughout this dissertation as

[11] 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.

rhetorical devices. In service of arguing honestly, therefore, my designs — in the body of the dissertation and in Appendix C — are accompanied by footnoted explanations of my design rationale.

This dissertation understands archives, bibliographies, anthologies, and corpora to all be, 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[12]. 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. One challenge to studying these models is that they present a “moving target”: even a bibliography or anthology is subject to change through successive editions (not to mention their now-common digital supplements), and a digital database has the potential to

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

change daily. I follow Kath Bode's approach in *A World of Fiction*, in artificially "freezing" each resource for study, and presenting my analysis as a description of a snapshot in time.[13] 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.

1.4. Scope

All of the computational work in this dissertation aims to identify, in as minute detail as possible, all works printed in England between January 1 1789 and December 31 1799. This eleven-year "decade" was a turbulent one across the Channel, encompassing the whole of the French Revolution, from the Estates General in 1789 to Napoleon's coup in 1799.[14] 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

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

[14] 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.

the decade of *Rights of Man*, 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, which are no longer neatly periodized as mutually exclusive.

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, 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 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.

Some limiting factor was necessary to make this project feasible from a technical standpoint. I needed to define a small enough scope that I could attempt something like comprehensiveness within that scope. I also knew that for many of the databases I wanted to use, I would not be able to access their full records, but samples of up to roughly 50,000 records had been given to other scholars (based on other papers I saw people publish on the ESTC). I decided that roughly a decade would give me enough texts to be worth approaching in this way, but not more texts than I could handle. I narrowed my focus to England as a way to sidestep problems of metadata. The Ireland of the eighteenth century had an unstable and contested relationship with the United Kingdom: how is this addressed by metadata assigning countries to texts? Is something tagged as “UK” if the city is in the UK now, or if it was in the UK at the time? I assume now. I decided not to open this can of worms! “Future Work.” Narrowing to England also helps to reduce the number of texts being considered. I’m disappointed that this means excluding all those Scottish pirated editions. Additional Future Work could compare Scotland to England, or grapple with Ireland. Once I decided to pick roughly a decade, I picked the 1790s for a few reasons. I began in the eighteenth century as a Gothicist, and remain curious about how the Gothic might relate to its print context: choosing a decade that was important to the Gothic opened up the possibility that my ultimate findings would be Gothic-related. The 1790s were also just a generally exciting decade, as there was a massive expansion of print, and a massive cultural anxiety about the role of print in peoples’ lives. The nature of the project meant that I had to pick my decade before I really knew what to find in it: the heightened revolutionary stakes of this particular decade seemed to give me the best chance that the decade would have something interesting in it. It’s also an important decade for the eighteenth century texts which would later become canonical, namely, the origins of Romanticism.

1.4.1. Charlotte Smith

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

In addition to being a prolific and interesting writing who was prominent across multiple genres (of which there are many other authors), Charlotte Smith offers an interesting case study in a ‘successful’ recovery project. This allows me to ask: to what extend do research infrastructures ‘lag behind’ scholarly consensus? No eighteenth centuryist is now likely to say that Smith is irrelevant or unimportant to the period. In the infrastructure of literary canons as described by Guillory, she has certainly succeeded: she is given prominent space in all anthologies of Romantic literature; she regularly appears on introductory syllabi, including surveys of all British literature; there are scholarly editions, and seminars, and dissertations, conference panels, and every other sign that she is an important and valued writer. But what about digital infrastructures? Are they “up to date”? For many I would say, not really. The ESTC ecosystem is still strongly shaped by editorial decisions made at the time of microfilming, or at the time of indexing. On Wikipedia, Smith herself had a respectable wiki article, but none of her major works were covered until I began to create those articles myself over the course of this dissertation.[15]

Smith is also self-consciously navigating a series of questions about genre that interest me, namely, the fact that not everybody reads everything, or for the same reasons — she is not the same author from genre to genre. “Very few Smith scholars work actively on both the novels and the poetry, 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. Because the novel during the Romantic period is undergoing an extraordinary amount of change and innovation, as it moves closer to its modern form, editors of the novels (myself included) tend to focus on Smith’s techniques and innovations, her use of tropes and themes, her facility with genres and description. Conversely, because Romantic poetry in the Smithian tradition is so closely tied up with explorations of selfhood and subjectivity, memory and a personalized past, editors of the poetry tend to present it as reflective of a personalized state of mind, of ‘woman’s’ experience, treating its manifold themes and narratives as, finally, reducible to and manifested from Smith’s life. 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**?** Genre, it seems, carries a greater force in constructing our preconceptions of identity than has been recognized, and Smith is a case in point, a case we can crack by studying closely Smith’s style and techniques *across* genres.” (Labbe 5) In other words,

[15] One reason this dissertation does not engage with Wikipedia as an object of study, despite many interesting implications, is because I am too involved as a wiki editor to maintain an arms-length distance.

is poet-Smith different or separate from novelist-Smith of didactic-Smith? If so, does that tell us something about poetry or novels, aesthetically? About the marketing or consumption of poetry or novels? One of my core interests is grappling with heterogenous groups of texts, which are usually examined each in isolation, and trying to bring them together. This might be one of those things that people don't do because it's a bad idea, rather than because it's hard; maybe there aren't any meaningfully questions that can be posed about All Writing. But I still wonder: are these things as separate from each other as we assume? Are we *really* dealing with different Smiths, or do we just *expect to find* different Smiths?

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 wranglings over the inheritance worth nearly £36,000 soon arose and were not settled until almost forty years later. By 1783 Benjamin had already unlawfully squandered more than a third of this trust and, as a consequence, found himself first in deep debt and then in King's Bench Prison.” (Roberts). After the success of the *Elegiac Sonnets* allowed Smith to pay for her husband’s release from prison, Benjamin Smith fled to France to escape further creditors. Charlotte Smith moved between England and France over the next year and a half to negotiate his debts, and in 1785, the family was able to return to England. In 1787, after 22 years of marriage, Charlotte Smith legally separated from her husband, “an unusual step for a woman of her time” (Fry 7), and moved to a town near Chichester with her nine surviving children (of the twelve she had given birth to). However, despite this separation, Benjamin Smith retained a legal right to Charlotte Smith’s profits from her writing. Smith moved frequently after her separation, due to financial instability and declining health. “On 23 February 1806 Benjamin died in a debtors' prison and some money reverted to Charlotte Smith. By then she was far too ill to execute her favourite scheme, to settle on the shores of Lake 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.

1.4.2. Databases

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

The first three databases I examine will be no surprise to eighteenth-century scholars: ESTC, ECCO, and ECCO-TCP. Gale's Eighteenth Century Collections Online (ECCO), contains over 180,000 titles 1701-1800, of which 42,000 were printed in England between 1789 and 1799. ECCO is itself (mostly) a subset of the broader English Short Title Catalogue (ESTC), which contains more 460,000 texts 1473-1800, of which 51,965 were printed in England between 1789 and 1799 (indicating that nearly 10,000 titles in the decade appear in the ESTC but not ECCO).

The ESTC does not provide access to texts themselves: instead, it is an authoritative bibliographic catalogue, available as a searchable database. It is ECCO which provides texts: ECCO's 180,000 titles works are available as photographed facsimiles of the full text of each title. The facsimiles can be searched within ECCO's online interface; these searches examine a plaintext version of the facsimile pages that was generated by Optical Character Recognition (OCR), but this OCR text is not made directly available. As a result, the facsimiles may be read individually by scholars, but cannot form the basis for computational corpus analysis. A subset of ECCO's texts have been hand-prepared, as part of the Text Creation Partnership (TCP), to be easier to use in computational research. The resulting corpus of ECCO-TCP texts contains 2,231 titles, of which 466 were printed in England between 1789 and 1799. These titles are available as carefully edited texts encoded according to the Text Encoding Initiative (TEI) standard, which not only provides an accurate version of the text's words, but encodes substantial details regarding its context on the page. Most large scale distant reading of eighteenth-century literature relies on the ECCO-TCP corpus as its 'model' or 'sample' to represent the period. Accordingly, one of the tasks of this dissertation is to examine the makeup of this corpus, and how it differs both from other corpora and from print culture in the period itself. These three digital collections — ECCO, ESTC, and ECCO-TCP — are the primary digital resources for the period, which form the basis of most digital research. However, they represent only one approach toward the collection and presentation of digital texts, to which there are two broad kinds of alternatives. These large but meticulous collections occupy a middle space between, on the one hand, highly selective thematic collections, such as The Shelley-Godwin Archive, of which there are many, and the giants of indiscriminate textual accumulation, such as Google Books, of which there are few.

Smaller collections allow for more scholarly curation, but have corresponding limitations. Whereas the ‘main players’ of the the mega-archives can be easily enumerated, these specialized collections are numerous. Some will focus on particular kinds of texts, such as the Early Novels Database (2,041 novels 1700-1799) or Broadside Ballads Online (more than 30,000 broadside ballads). Others exhaustively index particular publications, such as *The Hampshire Chronicle* (1,950 references to fiction in issues from 1772-1829), the Index to the *Lady’s Magazine* (14,729 articles from 1770 to 1818), or the Novels Reviewed Database (1,836 reviews from *The Critical Review* and *The Monthly Review*, 1790-1820). Feminist scholarship in particular has seen the creation of resources like the Orlando Project, the Chawton House library Novels Online, Northeastern University’s Women Writers Online and UC Davis’s British Women Romantic Poets. The virtue of these collections is that they achieve even greater accuracy and comprehensiveness within their defined scope. The Shelley-Godwin Archive, for example, can reasonably aspire to digitize *every* known manuscript of Percy Bysshe Shelley, Mary Wollstonecraft Shelley, William Godwin, and Mary Wollstonecraft, and to provide these manuscripts in hand encoded plaintext transcripts. However, as is inevitable, these specialized archives have the vices of their virtues: their specialized focus allows them to adapt precisely to their materials, and their idiosyncratic data structures can rarely be combined with other resources. The William Blake Archive, for example, benefits enormously from designing its archive around the unique images of each page of each copy of each of Blake’s works. But because this approach is so well suited to Blake, it cannot be applied beyond Blake. Even if the archive’s resources were available for download, they could not be directly compared to materials from another source which does not record its information at such a minute level of detail. As a result, although a great deal of excellent digital scholarship is contained in specialized micro archives, I do not examine them further in this dissertation.

Instead, I look at a set of larger archives of more contested “scholarly” status: Google Books, Project Gutenberg, and HathiTrust. Google Books may be the most infamous database of books. In a scholarly context, one hesitates even to designate this as an “archive,” particularly in the same breath as resources like ECCO: books of all kinds are scanned indiscriminately with only the bare minimum of roughly accurate metadata collected about them. These rapidly scanned books are prone to unpredictable errors, including inaccurate dates, misspellings, duplicate copies, and inaccurate subject classifications — 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.

Also in the category of smaller and specialized archives is Project Gutenberg. Project Gutenberg makes no claims to scholarly reliability but nonetheless underlies a not-significant amount of scholarly work[16] — its cultural capital as a resource lags far behind its use and utility. Project Gutenberg is easily conceived of as a haphazard, ‘unscholarly’ source for materials, but unlike Google Books, Project Gutenberg actually does have selection criteria. Project Gutenberg will only collect public domain works which contemporary audiences might be interested in reading for pleasure. It narrows the field substantially to exclude works which have either ceased to be broadly interesting (as in the case of most forgotten fiction), or which were never

[16] I have heard it quipped more than once in DH gatherings that you always think you’re going to get your texts from somewhere else, but Project Gutenberg is where you’ll actually get them.

particularly interesting (as in the case of almanacs and tax codes). Project Gutenberg includes 57,796 texts: far more than specialized scholarly archives like the Early Novels Database or the Shelley-Godwin Archive, but nonetheless an order of magnitude fewer than its more voracious potential competitors. And, like smaller specialized scholarly archives, Project Gutenberg has tailored its holdings to make it easy for readers to read, and quite difficult for its collection to be applied to any other use. By tailoring the structure of the archive itself to its specific materials, these collections are able to thoughtfully achieve their aims — but they also make it correspondingly difficult for users to achieve their own, different aims.

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.[17] These tools

[17] For example, it might be able to acquire a text document with all of the words of a novel, but sorted

provide a unique solution to real barriers for digital scholars of contemporary literature: although copyright law would make it prohibitively expensive or even impossible to build corpora of post-1920s literature, HathiTrust's mediated access to these texts enables corpus analysis. Through its collection, HathiTrust provides a hodgepodge of texts, of often unverifiable provenance and accuracy, selected largely by happenstance and convenience in a quest to contain all printed books. Through its tools, however, and through its institutional affiliations, HathiTrust has acquired a cultural capital among scholars which Google Books still lacks.

HathiTrust's success in acquiring scholarly capital stands in interesting contrast with Project Gutenberg's continued lack of cachet. Project Gutenberg is used in research with similar frequency to Google Books' n-gram tool,[18] 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.

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.

[18] 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.

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

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

1.5. Dissertation Map

Chapter two describes in more detail the databases to be studied, and examines Charlotte Smith and the ways that her writing is made accessible today. The specific experimentation undertaken in chapter two tests the basic assumptions and methods of my project. I begin with a the histories of the ESTC, ECCO, ECCO-TCP, Project Gutenberg, Google Books, and HathiTrust: highlighting the chronological relationships between these resources can explain each database's scope and technical implementation. Each new resource must contend with the possibility of either competing or collaborating with those which have come before. Examining materials like the meeting minutes and internal communications of these resources' early histories will show how those which currently enjoy the lowest reputation among scholars — Project Gutenberg and Google Books — defined their initial scope around an explicit rejection of scholarly norms. After establishing the history of each resource's development, I describe its current digital infrastructure, through the lens of critical algorithm studies. This begins with basic questions: what file formats does it use? What kinds of metadata, what ontologies? How does it make its materials available for use? Through close reading and comparison of these details, I articulate each database's implicit construction of what a text is and what it is for.

Having established these databases as objects of study, I identify what subset of Smith's works each corpus contains, as a concrete example to compare their holdings overall. Smith's *Elegiac Sonnets,* for example, are not included in the ECCO-TCP corpus (which is the one most often used for text mining research) — only *Celestina* and *The Emigrants* are included. Why these two texts? And what text mining research based on ECCO-TCP might have found slightly different answers if Smith's sonnets had been included? As a related test of comparison between databases, for each database which provides access to the actual text of Smith's works, I compare the textual similarity of *Celestina* and *The Emigrants.* What editorial choices are being made? How *much* worse is the OCR text than the transcribed text? Another key concept I will explore through Smith is the role of reprints. HathiTrust, for example, includes multiple editions of *Elegiac Sonnets*. How reliable and effective are its distinctions between editions? How do the databases I examine handle multiple editions of a single work? I am particularly interested in how reprints can be incorporated into our understanding of what literature is "of" a particular decade: what does it mean to think of *Elegiac Sonnets*, initially printed in the 1780s, as "1790s literature"? Finally, having surveyed my six databases with the help of Smith, I discuss the multiple "Smiths" which emerge, and what it means to attempt to unify her disparate works.

In chapter three, I re-examine my core databases, but no longer with Smith as a focalizing lens. Instead, I undertake computational assessment and comparison of the databases' contents. My research examines the authorship and subject matter (broadly construed) of all titles printed in England between 1789 and 1799 which are included in each database. I calculate the proportion of the titles in each resource that are attributed to men, to women, or are left unsigned. My naive hypothesis is that, as each resource demanded a greater investment of scholarly effort in each text, women and unsigned writers will grow increasingly underrepresented, so that the ECCO-TCP corpus will have substantially different demographics than the ESTC. Using the titles of these works and a topic modelling tool which I have built, I also roughly identify the subject matter of each title, categorizing works into broad genres such as drama, poetry, Romance, History, or sermons. Although the topic modelling tool is able to cluster what it sees as "similar" titles, individual interpretation is required to make these clusters meaningful. A substantial portion of chapter three is dedicated to discussing how scholars apply genre categories retrospectively to clusters of texts, how publishers sought to advertise their texts to particular audiences, and how the categories I develop ought to be understood in the context of existing eighteenth century scholarship on print genres.[19] Using my resulting genre classifications, I am then able to compare these four resources to each other, and to existing scholarly work on the print production of the 1790s. For example, I compare each resource's holdings to the statistics on the English novel included in Garside, Raven and Schöwerling's *Bibliographical Survey of Prose Fiction*. In examining genres, I anticipate discovering a preference, in the more specialized resources, for more "literary" forms of writing.

[19] I am particularly excited to explore "false advertising" in titles!

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

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

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

Chapter One

This chapter will describe the primary object of study of this dissertation, namely, contemporary digital databases with substantial holdings of 1790s literature. These databases require substantial description, preliminary to analysis, since they have conventionally been treated as tools for accessing objects of study, rather than objects of study in themselves. Figure 1 shows a venn diagram of the approximate relative scale, and overlap in holdings, of the eight databases in three ecosystems which I will explore. In blue is the purely academic ecosystem: the English Short Title Catalogue (ESTC), Eighteenth Century Collections Online (ECCO), and the ECCO Text Creation Partnership (ECCO-TCP). In pink is the commercial Google-backed ecosystem: Google Books, Google Ngrams, and HathiTrust. And in green is a crowdsourced ecosystem: Project Gutenberg. The division of databases into these ecosystems represents my own analysis of the institutional processes and selection principles which have shaped them. Databases within a shared ecosystem may or may not be interoperable, but they made their initial textual selections with a similar logic, and make make their holdings available for a similar imagined audience. An immediate difference between them, for example, is that the commercial ecosystem treats textual holdings like a trade secret; my discussion of Google Books and Google Ngrams will be primarily a discussion of barriers and lacunae. Google Books and Google Ngrams are of crucial importance, however, for understanding HathiTrust, whose presence in my diagram of the commercial ecosystem runs counter to its current highly academic reputation. This chapter will begin with a discussion of the theoretical stakes involved in taking a system of databases as one’s object of study. To illustrate these states, I will examine the current scope and data structure of each database, and the holdings of works by Charlotte Smith that each contains.

Then I will present a chronological history of these nine databases. The chapter concludes by returning to this map of database ecosystems, to examine the role of commercial forces in defining each of the three clusters. As I will show, these resources are different in their core assumptions — so different it can even be difficult to bring them into simple comparison with each other — as a result of three different strategies of response to the economic forces of the emerging internet.

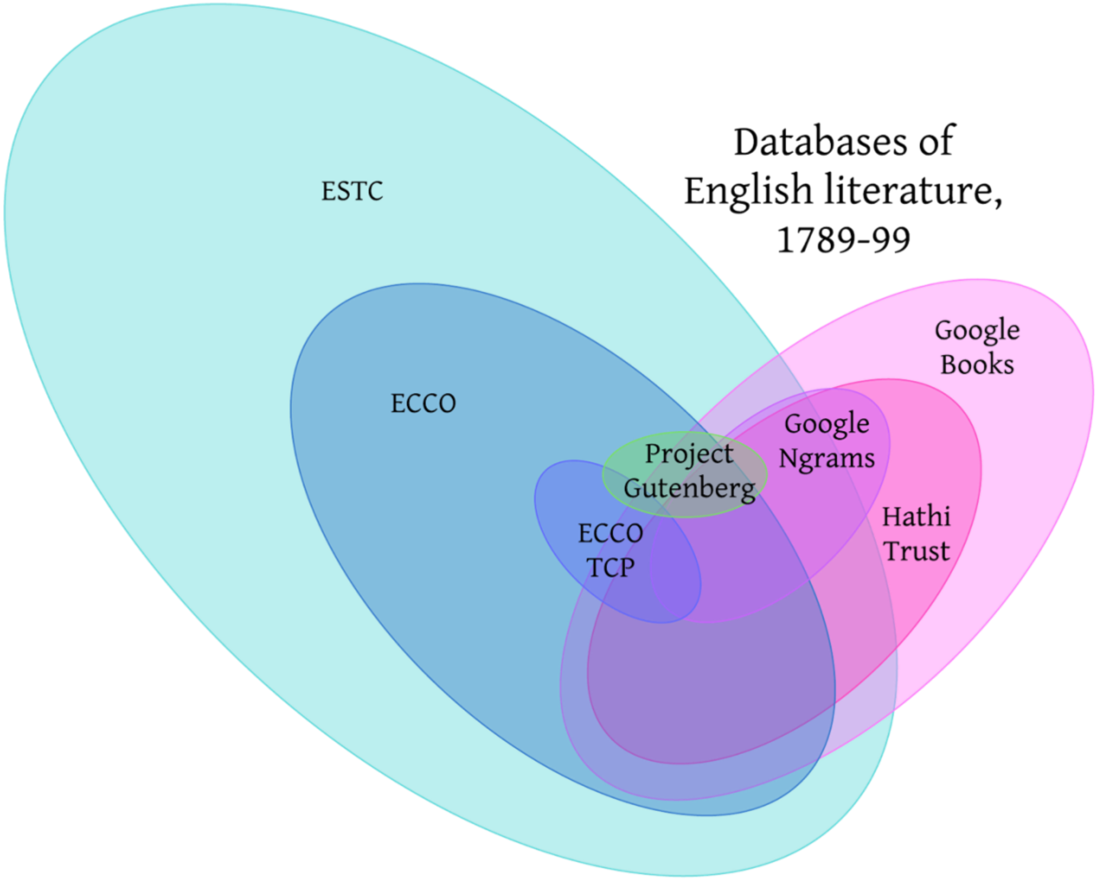

Figure 2: A hand-drawn venn diagram of the relative scale and overlapping holdings of databases containing 1790s literature. The orientation and colour-coding of the ovals groups them into three ecosystems: the academic databases in blue pointing left, the commercial databases in pink pointing right, and the crowdsourced databases in green oriented horizontally.

2.1. Why Write a History of a Database?

It is a common critique of Digital Humanities research that computational methods, or computational evidence, go hand in hand with claims of scientific, empirical, objective truth. Johanna Drucker's influential "Humanities Approaches to Graphical Display," for example, suggests that even by drawing a bar chart or line graph like those which appear in scientific journal articles, a humanities scholar goes astray. Drucker introduces a valuable contrast between "data" (an essentially imaginary phenomenon, of unmediated true information which a realist, empiricist researcher can accept as "given") and her term "capta" (the true form of all data, partial and ambiguous information, "captured" and constructed for a particular researcher's purpose). Drucker's distinction between data and capta is widely cited, especially by DH researchers who then go on to refer to their capta as data throughout. Drucker's core argument has, essentially, two parts:

> ...the basic categories of supposedly quantitative information, the fundamental parameters of chart production, are already interpreted expressions. But they do not present themselves as categories of interpretation, riven with ambiguity and uncertainty, because of the representational force of the visualization as a "picture" of "data". (12)

The first claim, that "data" is always already interpreted, is one that continues to animate and trouble DH scholarship. The second claim, that the primary site at which capta masquerade as data is through visualization, has received less uptake. In chapter two, as I produce visualizations of the role of women's writing in 1790s literature, I will return to Drucker and the role of data visualization in humanist research. It is certainly possible for visualizations to disguise the constructed nature of the underlying information, but, as Katherine Bode has recently shown in detail, the illusion of unmediated truth in computational research runs deeper than visual display.

Bode's article "The Equivalence of 'Close' and 'Distant' Reading" sets out to explain how distant reading, as a particularly headline-grabbing subset of DH, became so firmly associated, especially in the minds of its critics, with “the view of distant reading as enabling direct and objective access to a comprehensive literary-historical record.” (78) Bode contends that the exaggerated vision of distant readers who finally discover literary truths without getting bogged down by the pesky details of actual books — oracles who deliver scientific objectivity to rescue literature from its messiness — is not a vision wholly fabricated by media outlets and naysayers, but one fostered by the most prominent distant readers’ unacknowledged New Critical roots. Bode’s extended analysis of Franco Moretti and Matthew Jockers’ work contends that they “take the core premise of the New Criticism – that the text is the source of all meaning – to an extreme conclusion” (91) by relying on the accumulated contents of works of literature to explain the history of literary production. In the “Slaughterhouse of Literature” chapter of *Distant Reading*, for example, Moretti asks why Arthur Conan Doyle has become the canonized figurehead of early detective fiction. The mechanism for canonization that he proposes is that readers chose to continue reading Doyle over the years, in preference to other mystery writers, thus keeping Doyle “alive”; they made their selection based on formal features of Doyle’s stories; and specifically, Moretti suggests, they selected for the formal feature of “decodable clues.” Bode takes issue with this line of reasoning from its very first step: “Moretti takes as transparently true the idea that authors who have a canonical status in the present were selected from the time of first publication,” an assumption immediately countered by, for example, the relatively small readerships for five of the “big six” Romantic poets ([20]90). Then, Moretti’s assumption relies on taking a historical, social process (cultural capital accruing to “canonical” works in different

[20] Byron being, of course, the exception. Bode turns to William St. Clair to support this claim, and *The Reading Nation in the Romantic Period*. Cambridge: Cambridge University Press, 2014.

periods), and translating it into a purely formal, textual phenomenon to be observed directly within the work itself. Bode's final critique is about New Criticism directly, but applies equally well to this version of distant reading:

> …the assumption that literary works are texts, and that texts are single, stable, and self-evident entities, dismisses the documentary record's multiplicity, and with it the critical contributions of those disciplines (particularly bibliography and scholarly editing) that are dedicated to investigating that multiplicity. (92)

What Bode reveals, then, is that the monolithic concept of "data" can (as Drucker says) "[collapse] the critical distance between the phenomenal world and its interpretation" (1) with seductive ease regardless of whether visualization has been involved.

It is increasingly *de rigueur* for distant reading work to carefully discuss its underlying "corpus building." The terminology of "corpus building" goes perhaps halfway to the ideal humanistic relationship to knowledge which Drucker and Bode envision. By referring to the process as *building*, and discussing it as it own methodological step, scholars acknowledge their active intervention in creating the information they will then analyze. In Ted Underwood's monograph *Distant Horizons*, for example, he models many different kinds of literature to explore how texts might be distinguished — fiction vs non-fiction; popular vs literary; "genre fiction" in its many forms — and each model is first described in terms of how the definition of each group has been translated into a method of textual selection. All of this explicit discussion of corpus building serves to temper the strength of his arguments, cutting off the impossible illusion of "direct and objective access" (Bode "Equivalence" 78).

What is "out of scope" for almost all discussions of corpus building, even when framed as textual selection or sampling or modelling, is how the scholars go from the list of texts they wish to examine, to the digital surrogates for those texts. To be more explicit: underneath many a carefully-curated well-theorized corpus is, usually, a database. Again Bode is unusual, because she describes the creation, maintenance, lacunae, and even the irregular funding of the *Trove* database underlying her own corpora. More common is an approach like Underwood's, where methods sections dead-end in HathiTrust or, perhaps, the British Library. For essentially practical reasons, moreover, researchers often work within just one database ecosystem. However, "the suggestion implicit in databases such as EEBO and ECCO that this deluge of texts they make available is complete, constituting a form of universal coverage, leads the average users, especially the average students, to be lulled into a false sense that they now have access to "everything." This in turn reinforces a belief that if you don't find something—an author, a text, and a document—in these massive digital repositories, it did not exist" (Ezell 9)

This, then, is why the rest of this chapter will tell a long and perhaps even excruciatingly detailed history of databases: because the implicit arguments of textual selection must be made explicit, and a database, too, carries an implicit argument in its textual selection.

2.2. Eighteenth-Century Databases Today

Before delving into the more than fifty years of interlocking institutional histories which have created the current landscape of digital databases in eighteenth century studies, it is worth pausing to observe carefully what that current landscape actually looks like. One way to gain our bearings with an overwhelming quantity of information is to particularize it into anecdote: accordingly, our first look at these resources will be with the guiding figure of Charlotte Smith. Which works by Smith are in which database, why, and how? For the purposes of this chapter, I examine Smith's works which fall outside this dissertation's decade of interest. As Table 1 shows, Smith's publishing career began in 1784 and continued until her death in 1806; when I refer to Smith's "full" output, I consider all 47 editions of her works published in her lifetime or in the year immediately following her death. Her 1790s output (that is, the editions published 1789-99) consists of 30 of those editions. I have slightly expanded my chronological focus in part because some of the most interesting exclusions occur earlier and later in Smith's publishing career, such as the first edition of her immensely influential *Elegiac Sonnets* (1784), which is listed in the ESTC but not available in facsimile anywhere, or the publications in the last years of her life, which are excluded from the chronological focus of most resources but can still appear in HathiTrust. Of particular interest is the fact that *Beachy Head*, which is now one of Smith's most frequently anthologized and taught poems, does not appear in a single digital database. None of these inclusions or exclusions represent an agenda against (or for) Smith, or indeed an interpretive choice at all, but they nonetheless shape the disciplinary infrastructure.

year	**title**	**ed**	**ESTC**	**ECCO**	**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**	**3rd 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 3 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)[21]. 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.[22] 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

[21] 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."

[22] 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 copy, but the HathiTrust images are “Google-digitized” from the New York Public Library. Google’s rapacious book-scanning, evidently, was not as thorough as ECCO’s sustained scholarly project.

The smallest subset of all of these texts is the ECCO-TCP holding of just two titles: the second edition of *Celestina* (1791), and the first edition of *The Emigrants* (1793). Both titles appear in all larger databases, including HathiTrust (though, as I will discuss, they arrive in HathiTrust from a different source). *The Emigrants* is included in ECCO-TCP as one file, based on the ECCO facsimile of an original from the Huntington Library. *Celestina* is included as four files, one for each of four volumes, based on the ECCO facsimile of an original from the British Library. Both works were first reproduced in the microfilm version produced 1982-2002 in by Research Publications,[23] 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

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

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.

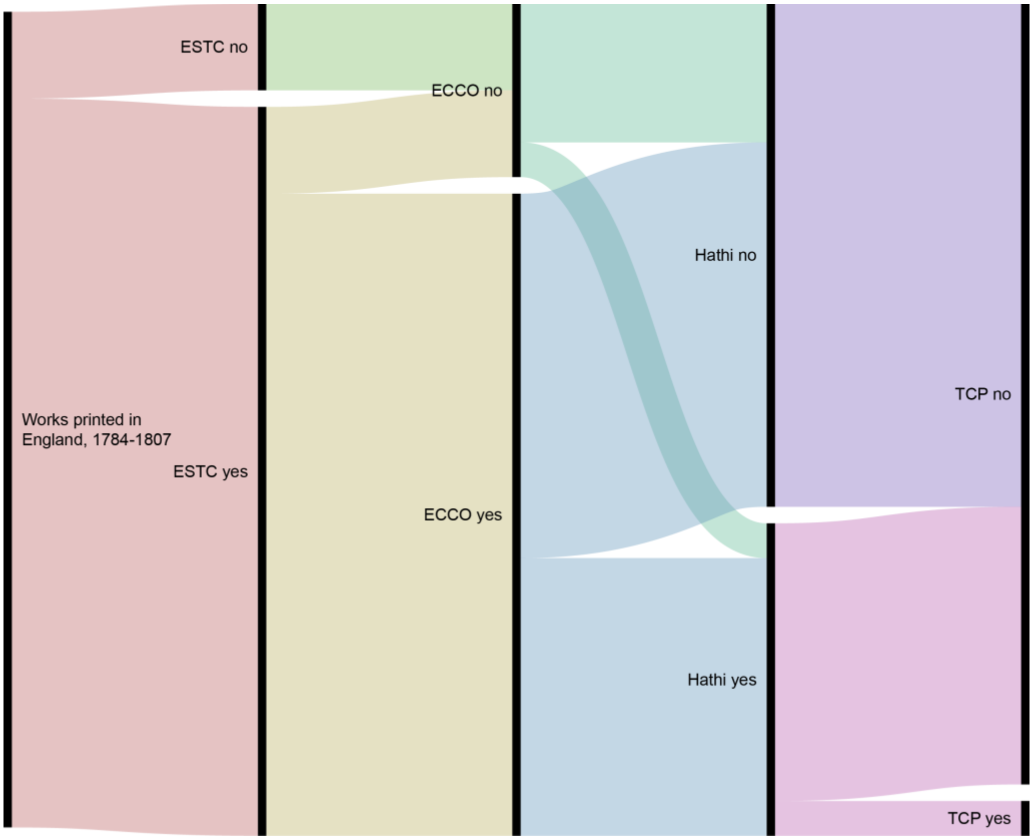

Figure 3: An alluvial chart, showing the winnowing down of Smith’s works from database to database. Of the 47 editions printed in England between 1784 and 1807, 42 are included in the ESTC, and 5 do not appear in the ESTC because they were printed after 1800 and thus fall outside its purview. ECCO contains 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.).

The Smiths outside of these datbases also differ. Only one of Charlotte Smith’s works is available in Project Gutenberg: *Emmeline, the Orphan of the Castle* (first published 1788). Searching the ESTC for records which both have “Toronto” in the library name and “Charlotte Turner” in the author name turns up two records: volume one of *Rural Walks* (1795) and *Minor Morals* (1798), both held at the Toronto public library. The Toronto Public Library 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 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.

2.3. History of Databases

Although each individual database has attracted some discussion, they have not been discussed together as an interlocking system. Instead, each database (or cluster of related databases) attracts discussion within its own conventional sphere. Here, I will discuss them together, organized chronologically by decade. A chronological organization makes it clear how many different organizations are in fact responding to shared historical conditions, or even to each others' development, as they make strategic decisions over time. A chronological organization also contrasts with essentially teleological descriptions of individual resources, which tend to work backward from a current state to present a clean narrative of how that current state was discovered to be ideal. By refusing to gloss over dead ends, periods of stagnation, and other oddities, we can better understand the current state of contemporary databases as the outcome of historically contingent processes which might have turned out differently, rather than accepting them uncuriously as inevitabilities.

1918	Pollard first proposes a "short-title handlist"
1926	Pollard and Redgrave Short-Title Catalogue for 1476–1640
1938	Eugene B. Power founds University Microfilms
1945	Wing starts collecting his STC, 1641–1700
1951	Donald Wing's catalogue for 1641–1700, first edition
1971	First text in what would be Project Gutenberg. Over the next twenty years, Michael Hart personally keyed the first hundred books.
1972	Beginning of second ed of Wing STC, 1641–1700
1976	Proposal for Eighteenth Century Short Title Catalogue, British Library and the American Society for Eighteenth Century Studies
1976	Second edition, vol 1, of Wing's STC
1976	Beginning of second ed of Pollard & Redgrave STC, 1475-1640
1977	ESTC pilot begun at British Library, directed by Robin Alston
1979	ESTC: Libraries from USA, Germany, and Australia began contributing to ESTC
1980	ESTC database available via British Library BLAISE [British Library Automated Information SErvice]
1981	Research Publications, Inc begins microfilming books
1981	ESTC database available via US Research Libraries Group RLIN [Research Libraries Information Network] system
1983	ESTC catalogue of BL holdings and indexes published in microform
1983	*Eighteenth Century Collection* microfilm produced by Research Publications, Inc
1985	ESTC online databases in RLIN and BLAISE upgraded to allow dynamic updates to a single shared file
1986	Second edition, vol 2, of Wing's STC
1987	ESTC expanded scope to add all print prior to 1700, changing its name to the English Short Title Catalogue. Information from Wing and STC is added to ESTC.
1987	Michael Hart recruits first Project Gutenberg volunteers
1989	Project Gutenberg completes its tenth book, the King James Bible
1991	End of second edition of Pollard & Redgrave STC, 1475-1640
1991	Exhaustive index to Wing's STC; end of Bibliographical Society support
1992	ESTC expanded scope to add serials
1994	ESTC made pre-1700 records available
1994	Project Gutenberg completes its 100th book, the Complete Works of Shakespeare

1994	Project Gutenberg's first website is developed by volunteer Pietro Di Miceli
1997	Project Gutenberg publishes its 1000th book, La Divina Commedia di Dante, in Italian
1998	ESTC second edition released on CD-ROM
1998	Conclusion of second ed of Wing STC
1998	Beginnings of EEBO: University Microfilms (now ProQuest) began to make available digitised copies of its microfilms across the Internet to subscribing institutions
199 9	ESTC assumed official responsibility for receiving new Wing STC data
1999	TCP began
2000	Charles Franks launches Distributed Proofreaders, to proof Project Gutenberg texts
2003	ESTC third edition released on CD-ROM
2003	Project Gutenberg 600 "best" ebooks released on CD-ROM, followed by 10,000 on DVD
2003	Beginning of ECCO: Thomson Gale (now Gale Cengage Learning) made digital copies of Eighteenth Century Collection microfilms available to subscribers online
2004	Google Print is announced
2005	TCP begins encoding ECCO texts
2006	ESTC available to search free online; ESTC begins transcribing full title and imprints
2007	Project Gutenberg DVD released with 17,000 items
2008	Project Gutenberg publishes its 25,000th book
2008	HathiTrust founded, by 12-university Committee on Institutional Cooperation and 11-library University of California Libraries
2009	EEBO-TCP Phase I complete: produced 25,000 books; beginning of Phase II
2010	Project Gutenberg DVD released with 30,000 items
2011	40,000 books in Project Gutenberg
2015	EEBO-TCP Phase I books released to the general public
2017	Project Gutenberg discontinues free mailing of CDs and DVDs, though the files remain available for people to burn their own copies at home
2021	EEBO-TCP Phase II books scheduled to be released to the general public

Table 1: A chronological history of major events in the development of six databases: the English Short-Title Catalogue (ESTC), Eighteenth Century Collections Online (ECCO), the Text Creation Partnership (TCP), Project Gutenberg, Google Books, and HathiTrust. Also included are events in the development of related resources, such as Early English Books Online (EEBO).

2.3.1. 1970s

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 published its first volume in 1976. 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.

The English Short Title Catalogue 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 and Donald Wing's catalogue for 1641–1700. 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. 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 lengthly 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).

Project Gutenberg began in 1971 with one individual, Michael Hart, who did not begin with a specific project vision in mind. From the beginning, then, Project Gutenberg was not goal-oriented in the same way as the other resources under discussion. By this I mean that Project Gutenberg orients itself toward goals of a fundamentally different kind than the goals which structure other textual archives, not that it has no goal. Project Gutenberg is, in general, subject to being dismissed as unserious or lacking rigorous standards, but I argue that these dismissals come from a failure to recognize and respect the real goals, seriousness, and standards which drive the project. In the case of the project's founding, that goal was not, as in the case of the other databases under discussion, to provide a particular kind of access to a particular kind of texts. Instead, the goal of Project Gutenberg was born from a moment of happenstance and nepotism by which Hart, a student at the time, was donated $100,000,000 of computer time on the Xerox Sigma V mainframe at the Materials Research Lab at the University of Illinois. This mainframe was one of the first fifteen nodes on the early ARPANet, the precursor to the modern internet. As Hart described it, he "decided there was nothing he could do, in the way of 'normal computing,' that would repay the huge value of the computer time he had been given ... so he had to create $100,000,000 worth of value in some other manner" ("History and Philosophy"). Rather presciently for 1971, Hart concluded that the greatest value computing would offer was the storage, searching, and retrieval of other materials. He therefore typed up and distributed the Declaration of Independence.[24] This became the first text of what would eventually become Project Gutenberg. It might even be considered the first ebook (according to Lebert 2008). Project Gutenberg was certainly "the first information provider on the internet and is the oldest

[24] Later discussion of Project Gutenberg will more fully explore the implications of this textual selection.

digital library" (Lebert). "During the fist twenty years, Michael Hart himself keyed in the first hundred books, with the occasional help of others from time to time." (Lebert) "when we started, the files had to be very small as a normal 300 page book took one meg of space which no one in 1971 could be expected to have (in general). So doing the U.S. Declaration of Independence (only 5K) seemed the best place to start. This was followed by the Bill of Rights — then the whole US Constitution, as space was getting large (at least by the standards of 1973). Then came the Bible, as individual books of the Bible were not that large, then Shakespeare (a play at a time), and then into general work in the areas of light and heavy literature and references." (Hart "History and Philosophy") "That edition of Shakespeare was never released, due to copyright changes. If Shakespeare's works belong to the public domain, the comments and notes may be copyrighted, depending on the publication date. But other editions belonging to the public domain were posted a few years later." (Lebert)

2.3.2. 1980s

In 1980, the ESTC began to go online. "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 LibraryAutomated 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 mean time “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.”(Norman)

The book facsimiles which would become Eighteenth Century Collections Online (ECCO) began as in 1983, when the company Research Publications, Inc began to produce its *Eighteenth Century Collection* microfilm. Research Publications was a newly-founded for-profit company, which was founded in 1981. They and their rival, University Microfilms, produced many of the facsimile images in contemporary databases. Today, the former Research Publications, Inc is part of Gale Cengage, and University Microfilms is part of ProQuest; as ECCO's history continues, the private company that owns the microfilms will change many times. "these documents were imaged in the late 1970s, transformed into microfilm during the 1980s" (Christy et al. 1) The second volume of the second edition of the STC was published in 1986.

1989 is also when Project Gutenberg began to make use of OCR to generate base texts which were then proofread, rather than having the text typed from scratch. (Paywalled WSJ article) "In August 1989, Project Gutenberg completed its 10th book, The King James Bible, that was first published in 1611, with the standard text dated 1769" (Lebert) "Then, through being involved in the University of Illinois PC User Group and with assistance from Mark Zinzow, a programmer at the school, Hart was able to recruit volunteers and set up an infrastructure of mirror sites and mailing lists for the project. With this the project was able to grow much more rapidly." (History-Computer)

2.3.3. 1990s

In the 1990s, the ESTC began to expand its scope. "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). The ESTC continued to research new entries and improve existing ones, releasing a second edition of the file on CD-ROM in 1998.

The second edition of the STC completed its publication in 1991, with a set of exhaustive indexes to its material. 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). "EEBO's relationship with the original STC and Wing is straightforward and clear; EEBO's relationship with electronic ESTC, on the other hand, is less well-known.20 A series of agreements made between ESTC and University Microfilms/ProQuest between 1989 and 1997 allowed EEBO to draw directly on ESTC's existing bibliographical data. Consequently, / every search run on EEBO (with some exceptions) relies, in a fundamental sense, on bibliographical information originally supplied by ESTC – but not in the form that one might expect. First, EEBO heavily edited ESTC's data for its own purposes: certain categories of data were removed (e.g. collations, Stationers' Register entrances), some information was amended (e.g. subject headings), and some was added (e.g. microfilm- specific details). Second, there is no formal mechanism for synchronising the data between the two resources. Occasionally, snapshots of data are sent by EEBO to ESTC but there is no guarantee that a correction or revision made to an ESTC entry will be replicated in the corresponding EEBO entry or vice versa: neither ESTC nor EEBO will necessarily know when the other has made a correction. As both resources continue to amend and expand their bibliographical data for their own purposes, there is an increasing likelihood of significant discrepancy between the two resources. Finally, although EEBO continues to microfilm and digitise, there is no absolute one-to-one correspondence between the pre-1701 entries in ESTC and the materials on EEBO; there are – and will always be – items on ESTC not available on EEBO." (Gadd 685-6) By 1997, Research Publications, Inc had become Primary Source Media. In September 1998, "the Thomson Corporation [merged] three of its electronic and reference publishing subsidiaries—Gale Research, Information Access Company (IAC), and Primary Source Media—into a new company called The Gale Group."

Meanwhile, the 1990s saw Project Gutenberg's explosion of success, growing beyond a small group of hobbyists as the internet itself greatly expanded its reach. "When the internet became popular, in the mid-1990s, the project got a boost and an international dimension. Michael still typed and scanned in books, but now coordinated the work of dozens and then hundreds of volunteers in many countries." (Lebert) "In 1990, there were 250,000 internet users, and the standard was 360 K disks. In January 1991, Michael typed in Alice's Adventures in Wonderland, by Lewis Carroll (published in 1865). In July 1991, he typed in Peter Pan, by James M. Barrie (published in 1904)." (Lebert) "By the time Project Gutenberg got famous, the standard was 360K disks, so we did books such as Alice in Wonderland or Peter Pan because they could fit on one disk. Now 1.44 is the standard disk and ZIP is the standard compression; the practical filesize is about three million characters, more than long enough for the average book." (Hart "History and Philosophy") "Project Gutenberg gradually got into its stride, with the digitization of one book per month in 1991, two books per month in 1992, four books per month in 1993 and eight books per month in 1994. In January 1994, Project Gutenberg celebrated its 100th book by releasing The Complete Works of William Shakespeare." (Lebert) In 1994, Italian volunteer Pietro Di Miceli developed and administered the first Project Gutenberg website and started the development of the Project online Catalog. ("Credits")

2.3.4. 2000s

The 2000s are really the decade of Google Books. In 2002, "A small group of Googlers officially launches the secret "books" project. They begin talking to experts about the challenges ahead, starting with a simple but crucial question: how long would it take to digitally scan every book in the world? It turns out, oddly enough, that no one knows. In typical Google fashion, Larry Page decides to experiment on his own. In the office one day, he and Marissa Mayer, one of our first product managers, use a metronome to keep rhythm as they methodically turn the pages of a 300-page volume. It takes a full 40 minutes to reach the end. Inspired by the extraordinary digitization projects underway all around the world – the Library of Congress's American Memory project, Project Gutenberg, the Million Book Project and the Universal Library, to name only a few – the team embarks on a series of site visits to learn about how they work. As part of this fact-finding mission, Larry Page reaches out to the University of Michigan, his alma mater and a pioneer in library digitization efforts including JSTOR and Making of America. When he learns that the current estimate for scanning the university library's seven million volumes is 1,000 years, he tells university president Mary Sue Coleman he believes Google can help make it happen in six." ("Google Books History," 2016)

In 2003, the team works to develop a high-speed scanning process — "A team member travels to a charity book fair in Phoenix, Arizona, to acquire books for testing non-destructive scanning techniques. After countless rounds of experimentation, the team develops a scanning method that's much gentler than current common high-speed processes. … At the same time, the team's software engineers make progress toward resolving the tricky technical issues they encounter processing information from books that contain odd type sizes, unusual fonts or other unexpected peculiarities – in 430 different languages." ("Google Books History," 2016) In 2004, a formal partnership with Bodleian library begins, "to digitize the library's incomparable collection of more than one million 19th-century public domain books within three years." ("Google Books History," 2016) "In October, Larry and Sergey announce "Google Print" at the

Frankfurt Book Fair in Germany. The first publishers to join the program: Blackwell, Cambridge University Press, the University of Chicago Press, Houghton Mifflin, Hyperion, McGraw-Hill, Oxford University Press, Pearson, Penguin, Perseus, Princeton University Press, Springer, Taylor & Francis, Thomson Delmar and Warner Books. In December, we announce the beginning of the "Google Print" Library Project, made possible by partnerships with Harvard, the University of Michigan, the New York Public Library, Oxford and Stanford. The combined collections at these extraordinary libraries are estimated to exceed 15 million volumes." ("Google Books History," 2016)

The enormous and relentlessly physical labour underlying the project is almost impossible to imagine, despite the fact that it actually occurred: "Every weekday, semi trucks full of books would pull up at designated Google scanning centers. … The books were unloaded from the trucks onto the kind of carts you find in libraries and wheeled up to human operators sitting at one of a few dozen brightly lit scanning stations, arranged in rows about six to eight feet apart. … Each one could digitize books at a rate of 1,000 pages per hour. The book would lie in a specially designed motorized cradle that would adjust to the spine, locking it in place. Above, there was an array of lights and at least $1,000 worth of optics, including four cameras, two pointed at each half of the book, and a range-finding LIDAR that overlaid a three-dimensional laser grid on the book's surface to capture the curvature of the paper. … What made the system so efficient is that it left so much of the work to software. Rather than make sure that each page was aligned perfectly, and flattened, before taking a photo, which was a major source of delays in traditional book-scanning systems, cruder images of curved pages were fed to de-warping algorithms, which used the LIDAR data along with some clever mathematics to artificially bend the text back into straight lines. At its peak, the project involved about 50 full-time software engineers." (Somers)

In 2005, a controversy emerges about Google's "Library Project." "In keeping with our mission to organize the world's information and make it universally accessible and useful, we donate $3 million to the Library of Congress to help build the World Digital Library, which will provide online access to a collection of rare and unique items from all around the world. We also extend our pilot scanning program with the Library, which includes digitizing works of historical value from the Library of Congress Law Library. Google renames "Google Print" Google Books, which more accurately reflects how people use it. The team also responds to the controversy over the Library Project by engaging in public debate about its underlying principles." ("Google Books History," 2016) Undeterred, in 2006 there are several expansions to the project. "We launch a series of product enhancements to make Book Search more useful and easier to use. First, we expand access to the public domain works we've scanned by adding a download a PDF button to all out-of-copyright books. A few months later, we release a new browsing interface that makes it easier to browse and navigate Book Search. The new interface is also accompanied by new About this Book pages which use Google algorithms to populate pages with rich related content on a book -- initially, related books, selected pages and references from scholarly works. In the fall, four new libraries join the Library Project: the University of California, University Complutense of Madrid, the University of Wisconsin- Madison and the University of Virginia." ("Google Books History," 2016) In 2007: "Using the new UI as a launching point, we experiment with new ways for people to interact with books. Places in this Book: A mashup with Maps lets people browse books by locations mentioned in the text (later, we release an experimental KML layer for Google Earth that does the reverse -- the user picks a location, and we map books to it). Popular Passages: We create a new way to navigate between books, tracking the use of a single passage through a collection of books. My Library: We help people harness the power of Google search within their own personal book collections. Users begin to curate and share their personal libraries, reviews and ratings with others. New homepage (initially US only): We give people more jumping off points for exploring the books in our index. … we add a "View plain text" link to all out-of-copyright books. T.V. Raman explains how this opens the book to adaptive technologies such as screen readers and Braille display" ("Google Books History," 2016) "By December, the Book Search interface is available in over 35 languages, from Japanese to Czech to Finnish. Over 10,000 publishers and authors from 100+ countries are participating in the Book Search Partner Program." ("Google Books History," 2016) "In May, the Cantonal and University Library of Lausanne, and Ghent University Library join the Book Search program, adding a substantial amount of books in French, German, Flemish, Latin and other languages, and bringing the total number of European libraries partners to six. … Over 10,000 publishers and authors from 100+ countries are participating in the Book Search Partner Program. The Library Project expands to 28 partners, including seven international library partners: Oxford University (UK), University of Complutense of Madrid (Spain), the National Library of Catalonia (Spain), University Library of Lausanne (Switzerland), Ghent University (Belgium) and Keio University (Japan)."("Google Books History," 2016)

The "Google Books History" page ends with 2007, with the final words of "As we look to the year ahead, we continue to develop our technology and expand our partnerships with publishers and libraries all around the world. Stay tuned..." but the page spent more than a decade with no further updates.[25] 2008, the year that Google stops being quite so proud of the

[25] Some time between August 2019 and March 2020, the page was edited so that it no longer had a

rapid progress of Google Books, is also the year that they begin to face legal repercussions for their “scan first and ask questions later” approach to mass digitization. Somers has characterized this process as equivalent to the burning of the library of Alexandria: “When the most significant humanities project of our time was dismantled in court, the scholars, archivists, and librarians who’d had a hand in its undoing breathed a sigh of relief, for they believed, at the time, that they had narrowly averted disaster” (Somers). October 2008: A settlement was reached between the publishing industry and Google after two years of negotiation. Google agreed to compensate authors and publishers in exchange for the right to make millions of books available to the public. November 2008: Google reached the 7 million book mark for items scanned by Google and by their publishing partners. 1 million were in full preview mode and 1 million were fully viewable and downloadable public domain works. About five million were out of print. December 2008: Google announced the inclusion of magazines in Google Books. Titles include New York Magazine, Ebony, and Popular Mechanics. February 2009: Google launched a mobile version of Google Book Search, allowing iPhone and Android phone users to read over 1.5 million public domain works in the US (and over 500,000 outside the US) using a mobile browser. Instead of page images, the plain text of the book is displayed. May 2009: At the annual BookExpo convention in New York, Google signaled its intent to introduce a program that would enable publishers to sell digital versions of their newest books direct to consumers through Google. December 2009: A French court shut down the scanning of copyrighted books published in France, saying this violated copyright laws. It was the first major legal loss for the

year-by-year breakdown of milestones. Now, after briefly telling the anecdote about the 1996 origin of Google Books, the “history” says ”Fast forward to today: After more than a decade of evolution, innovation and strong partnerships, Google Books has helped to make more than 40 million books discoverable, in more than 400 languages. And we're not done -- not until all of the books in the world can be found by everyone, everywhere, at any time they need them.” (“Google Books History,” 2020) scanning project.

Academic institutions also spend the 2000s beginning make facsimiles available online. A third edition of the ESTC was published on CD-ROM 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). In 2003, Thomson Gale began making digital copies of the Eighteenth Century Collection microfilms available to subscribers online. The digital images were made from the microfilm masters, which were 400 dpi, and thus higher resolution than the microfilm copies, which Spedding reports were 300 dpi (440).

In 2000, Distributed Proofreaders was founded. Also “In 2000, a non-profit corporation, the Project Gutenberg Literary Archive Foundation, Inc. was chartered in Mississippi, United States to handle the project's legal needs. Donations to it are tax-deductible. Long-time Project Gutenberg volunteer Gregory Newby became the foundation's first CEO.” (Wikipedia) “A fast growth thanks to Distributed Proofreaders, a website launched in October 2000 by Charles Franks to share the proofreading of books between many volunteers. Volunteers choose one of the books listed on the site and proofread a given page. They don't have any quota to fulfill, but it is recommended they do a page per day if possible. It doesn't seem much, but with hundreds of volunteers it really adds up.” (Lebert) “In 2002 Distributed Proofreaders became part of Project Gutenberg.” (Hosch) “By 2009 roughly half of all Project Gutenberg books had been handled by using Distributed Proofreaders." (Hosch) "As of 2018, the 36,000+ DP-contributed books comprised almost two-thirds of the nearly 60,000 books in Project Gutenberg." (Wikipedia) At some point Carnegie Mellon University agreed to administer the project's finances. It's currently hosted by ibiblio at UNC Chapel Hill. "The number of electronic books rose from 1,000 (in August 1997) to 5,000 (in April 2002), 10,000 (in October 2003), 15,000 (in January 2005), 20,000 (in December 2006) and 25,000 (in April 2008). … The steady growth went on, with an average of 8 books per month in 1994, 16 books per month in 1995, and 32 books per month in 1996." (Lebert) "If 32 years were necessary to digitize the first 10,000 books, between July 1971 and October 2003, 3 years and 2 months were necessary to digitize the following 10,000 books, between October 2003 and December 2006." (Lebert) "In the first 11 weeks of 2004, Project Gutenberg added 313 new e-books. It took from 1971 to 1997 to produce the first 313 e-books—that's 11 weeks compared to about 26 years." (Hane) "In 2004 Project Gutenberg Europe and Distributed Proofreaders Europe were formed to facilitate the process of adding more non-English works." (Hosch) The books doubled again between 2006 and 2011, from 20,000 (2006) to 40,000 (2011). (Hosch) In 2004 Hart declared bombastically that "We want to grow the collection to 1 million free e-books and distribute them to 1 billion people for a total of 1 quadrillion e-books to be given away by the end of the year 2015." (Hane) – since the current collection is 'only' 60,000 ebooks, this ambitious scope was clearly not realized, but it reflected what was considered possible – or if not precisely possible, at least dreamable.

Project Gutenberg also explored making its materials available offline. In August 2003, Project Gutenberg created a CD containing approximately 600 of the "best" e-books from the collection. The CD was available for download as an ISO image. When users are unable to download the CD, they could request to have a copy sent to them, free of charge. In December

2003, a DVD was created containing nearly 10,000 items. At the time, this represented almost the entire collection. In early 2004, the DVD also became available by mail.

In July 2007, a new edition of the DVD was released containing over 17,000 books, and in April 2010, a dual-layer DVD was released, containing nearly 30,000 items. The majority of the DVDs, and all of the CDs mailed by the project, were recorded on recordable media by volunteers. However, the new dual layer DVDs were manufactured professionally, as it proved more economical than having volunteers burn them. As of October 2010, the project had mailed approximately 40,000 discs.

2.3.5. 2010s

The 2010s appear to have been a decade of stagnation for the ESTC, though not for lack of effort. 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. Reports on the outcomes of this work, however, are scarce.

The TCP began to set its eyes on the eighteenth century in 2004, as per their meeting minutes: "Jeff Moyer then updated the Board on its progress with the ECCO product which contains over 26 million pages and 155,000 volumes. To date, they have 60 customers including 6 Canadian and 11 international institutions. They have also done OCR for the ECCO product and are interested in how the TCP text will work and integrate with their OCR." ("Meeting Minutes 2004-10-21."). "In 2005 the project expanded to include Gale-Cengage's Eighteenth-Century Collections Online (as well as Evans Early American Imprints by Newsbank). However, while the EEBO-TCP project flourishes (with around 40,000 texts transcribed so far), the work on ECCO-TCP stagnated at around 2,000 texts. As well as the main partner institutions of Michigan and Oxford that oess to the eighteenth-century TCP texts, so I've listed them below, with a few comments." (Gregg n. pag.) By October 2005, "the ECCO-TCP project has commenced with the release of a demo, and TCP sponsored an ECCO selection task force in August. … Rich Foley reported that ECCO now covers 120 subscribers and a recent purchase from the JISC. The ECCO product has also sold well in Canada this past year. Rich also reported (relating to a question on Metadata) the release of a My Library product which is opening up access to their metadata at Gale and that he was interested in further doing case studies on how ECCO is used in research and in the classroom." ("Meeting Minutes 2005-10-20.") "In 2005, the TCP executive board and staff sought to expand the TCP model to other databases of historical books, namely, Gale Cengage's Eighteenth-Century Collections Online (ECCO) and Newsbank Readex's Evans Early American Imprints (Evans-TCP). These projects never received quite the support attracted by EEBO-TCP, and in the end produced only about 8,000 texts, compared to the 60,000 produced by the latter, with another few thousand on the way." (TCP "About") As early as 2003, the TCP executive board meeting minutes reported that "Michigan has made agreements with Gale and Readex to support conversion of subsets of the Eighteenth Century and Evans Early American materials which will allow us to create a cross-searchable corpus of important historical texts … The University of Michigan has reached agreements to create a subset of accurately keyed and encoded texts in conjunction with these projects, and aims to produce 6,000 early American and 10,000 18th century texts. In the near term, this will not affect production of EEBO texts because there is adequate capacity to expand beyond existing levels of production. In the long term, this will produce a large number of culturally significant texts, produced to a single standard, that are owned by the library community and complement the EEBO texts for these early historical periods." ("Meeting Minutes 2003-10-22.").

In 2006, ECCO-TCP was struggling compared to the other TCP products. "Rich Foley reported that ECCO is one of the biggest products at Gale with eighty to ninety ARLs subscribing as well as small institutions. He also said that a focus at Gale was to work on more tools to facilitate undergraduate teaching of their products. … Mark Sandler reported that the TCP budget shows mostly positive balances through 2007. The exception to that is ECCO but because it is still so early in the project, it seems likely that TCP will overcome those problems within the next few months. Nonetheless, the TCP project in EEBO, Evans, and ECCO face potential budget deficits in fiscal year 2008." At this time, the TCP began to think about the end of the project: "the TCP should set a date to close the partnership (likely around 2010 given current commitments," partly to address financial solvency. ("Meeting Minutes 2006-09-16"). In 2007, all three TCP project reported successful sales, though ECCO's news was the most vague: "Brandon Nordin also reported good news from Gale and along with Mary Sauer-Games announced that the EEBO and ECCO databases will now be cross-searchable so that users can go to either collection and find records from the other" ("Meeting Minutes 2007-10-30"). Nonetheless, "Evans-TCP and ECCO-TCP sales have historically (for a variety of reasons, chiefly the presence of OCR text in both projects) been weaker than anticipated" ("Meeting Minutes 2007-10-30"). And the end loomed nigh: "Currently finances are good through fiscal year 2008. EEBO-TCP is on target to complete 25,000 texts by the end of fiscal year 2008. Evans-TCP is likely, given current finances to complete around 6,000 texts. ECCO-TCP will complete around 1,300 texts. Therefore, the TCP, particularly in EEBO-TCP has been a success meeting most of its goals. Nonetheless, Evans-TCP and ECCO-TCP are still short of their goals of 6,000 and 10,000 texts respectively, and in fiscal year 2009, the TCP overall is facing a deficit of around $400,000 if it does not either reduce its current staff or bring in a large influx of money within the next six months" ("Meeting Minutes 2007-10-30"). These are the last meeting minutes available online.

However, despite the grim financial picture in the 2009 minutes, the project's goal outputs are initially rosy: "EEBO-TCP met its goal of producing 25,000 books in 2009 (thereafter known as "EEBO-TCP Phase 1"), and then undertook work on a second phase to convert the first edition of each remaining unique monographic work in EEBO—another 40,000 or so books, for a total of around 70,000, if all hopes were realized." (TCP "About") "Begun in 2009, Phase II both shrank and expanded the scope of EEBO TCP. Selection became more discriminating and focused more on English-language (and Welsh- and Gaelic-language) texts to the exclusion of French and Latin titles, and also set aside the serials (periodicals) as a fit project for another time. But within the constraints of English-language monographic titles, it aspired to something approaching comprehensive treatment: EEBO Phase II planned to convert each and every unique work in Early English Books Online (usually the first edition), or an estimated total of around 45,000 books on top of the 25,000 completed in Phase I. This was an ambitious, and always risky, goal. As it happened, enough institutions joined Phase II to fund the completion of about 40,000 titles, of which about 35,000 have been released to date, the remainder slowly working their way through the production pipeline. (TCP "EEBO") "As of 2019, the total number of books available in Phase II came to 34,963, with a further release of several thousand additional titles tentatively scheduled for later in the year. Short of an infusion of new funding, or the adoption of a new production model, this should bring the active work of the TCP to at least an interim conclusion." (TCP "EEBO") "Because of these greater challenges facing ECCO-TCP, it is perhaps better described as a proof of concept than as a completed project. With the support of more than 35 libraries, the TCP keyed, encoded, edited, and released 2,473 ECCO-TCP texts. A further tranche of 628 texts was keyed and encoded but never fully proofed or edited. The texts in this group remain useful for many purposes, however, and bring the total of ECCO-TCP texts to over 3,000. In cooperation with Gale Cengage, these texts have been made freely available to the public.” (TCP, “Eighteenth Century Collections Online (ECCO) TCP”)

In late 2019, Gale began allowing access to a new interface, the Gale Digital Scholar Lab, which dramatically changed the forms of access available for ECCO texts. It became possible not only to see the underlying OCR for texts, but to run pre-built text mining on it, and to download the OCR as text files. The only limit to downloading is that only 10,000 texts may be downloaded at a time, but as long as the desired corpus can be defined as “collections” in chunks of 10,000 or less, any number of files can be downloaded. In a particularly dramatic departure from ECCO’s past practice and current norms, I was told that there were also no restrictions on sharing the downloaded files, even though downloading them in the first place required a library subscription.

The decade was less good for Google Books, as the lawsuit against them finally drew to a close, though Google continued to scan at speed throughout. April 2010: Visual artists were not included in the previous lawsuit and settlement, are the plaintiff groups in another lawsuit, and say they intend to bring more than just Google Books under scrutiny. "The new class action," read the statement, "goes beyond Google's Library Project, and includes Google's other systematic and pervasive infringements of the rights of photographers, illustrators and other visual artists." May 2010: It was reported that Google would launch a digital book store called Google Editions. It would compete with Amazon, Barnes & Noble, Apple and other electronic book retailers with its own e-book store. Unlike others, Google Editions would be completely online and would not require a specific device (such as kindle, Nook, or iPad). June 2010: Google passed 12 million books scanned. August 2010: It was announced that Google intends to scan all known existing 129,864,880 books within a decade, amounting to over 4 billion digital pages and 2 trillion words in total. December 2010: Google launched the Ngram Viewer, which collects and graphs data on word usage across its book collection.

Despite the highs of 2010, however, in March 2011 a federal judge rejected the settlement reached between the publishing industry and Google.“At a time when the rest of Google was obsessed with making apps more “social”—Google Plus was released in 2011—Books was seen by those who worked on it as one of those projects from the old era, like Search itself, that made good on the company’s mission “to organize the world’s information and make it universally accessible and useful.” It was the first project that Google ever called a “moonshot.”” (Somers) March 2012: Google passed 20 million books scanned. March 2012: Google reached a settlement with publishers. November 2013: Ruling in Authors Guild v. Google, US District Judge Denny Chin sides with Google, citing fair use. The authors said they would appeal. October 2015: The appeals court sided with Google, declaring that Google did not violate copyright law. According to the New York Times, Google had by then scanned more than 25 million books. April 2016: The US Supreme Court declined to hear the Authors Guild's appeal, which means the lower court's decision stood, and Google would be allowed to scan library books and display snippets in search results without violating the law. Nonetheless, the project seemed to have lost its momentum, and the scanning process slowed down in American academic libraries. Three years later, in October 2019, Google celebrated 15 years of Google Books and provided the number of scanned books as more than 40 million titles. Google has been quite secretive regarding its plans on the future of the Google Books project. Scanning operations had been slowing down since at least 2012, as confirmed by the librarians at several of Google's partner institutions. At University of Wisconsin, the speed had reduced to less than half of what it was in 2006. However, the librarians have said that the dwindling pace could be a natural result of maturation of the project – initially stacks of books were entirely taken up for scanning whereas now Google only needed to consider the ones that have not been scanned already. Despite winning the decade-long litigation in 2017, The Atlantic has said that Google has "all but shut down its scanning operation." In April 2017, Wired reported that there were only a few Google employees working on the project, and new books were still being scanned, but at a significantly lower rate.

Some material is still being sent to Google Books for scanning, however. In November 2019, University of Colorado Boulder announced that their library would be partnering with Google for books to be scanned, with copies appearing both in Google Books and in HathiTrust. (“Increasing Access with Google Books”) “In total, the process is estimated to take two to four years to complete. With this estimation, Interim Director of Libraries Information Technology Michael Dulock approximated that if Google processed 200,000 books and each book was about 200 pages, this project will save the Libraries about $20 million. Another way to measure savings for the Libraries is with time. Dulock said that if Digital Media Services in the Libraries worked on this project full-time, with one staff member at 40 hours a week and five students at 20 hours a week, doing nothing else, it would take close to 100 years to complete.” (“Increasing Access with Google Books”) This press release puts a positive spin on the strange financial relationship between Google and the libraries it scans books from: as the enormous cost indicates, Google must be doing this scanning for its own financial benefit, and the library is therefore providing a service to Google – but the library is presented as the grateful, even grovelling, recipient of Google’s bounty simply for being permitted to share a limited copy of the value it has itself just given to Google.

The subject of library copies leads naturallyto HathiTrust, invented as a way to cobble together a Google Books alternative out of all those second-copies. ”HathiTrust was launched in 2008 by the 11 University of California libraries and the 12-university consortium known as the Committee on Institutional Cooperation (CIC), with key support provided by the University of Michigan and Indiana University.” (Karels) “As of today [October 13, 2008], HathiTrust contains more than 2 million volumes and approximately ¾ of a billion pages, about 16 percent of which are in the public domain. Public domain materials will be available for reading online. Materials protected by copyright, although not available for reading online, are given the full range of digital archiving services, thereby offering member libraries a reliable means to preserve their collections.” (HathiTrust, “Major Library Partners Launch HathiTrust Shared Digital Repository”) “When Google partnered with university libraries to scan their collections, it had agreed to give them each a copy of the scanning data, and in 2008 the HathiTrust began organizing and sharing those files. (It had to fend off the Authors Guild in court, too.) HathiTrust has 125 member organizations and institutions who “believe that we can better stewardresearch and cultural heritage by working together than alone or by leaving it to an organization like Google,” says Mike Furlough, the trust’s director.” (Van Helden) “The vast majority of those digitized books-around 95 percent, as of mid-2017- had originally been scanned as part of the Google Books project; the agreements that Google Books entered into with the libraries typically stipulated that Google had to provide the library with a digital copy of each book scanned from that library.” (Bauder) In 2010, just two years after its founding, ”HathiTrust is now jointly owned and operated by 52 institutions from the U.S. and Europe, all focused on a common goal

– to build an extraordinary digital library that preserves and provides access to the cultural record. The new members to HathiTrust include the Library of Congress, Stanford University, Arizona State University, Massachusetts Institute of Technology, and University of Madrid, HathiTrust's first international partner." (Karels) In October 2015, HathiTrust comprised over 13.7 million volumes, including 5.3 million of which were in the public domain in the United States. Which makes HathiTrust a strange hybrid: in some ways, merely a reskinned Google Books, with aesthetic tweaks to appear legitimate in scholarly eyes – in other ways, a genuinely effective way to reclaim noncommercial benefits from the activities of a megacommercial corporation. Certainly, in the absence of a direct profit motive underlying HathiTrust, it has more potential to eventually become the digital library of Alexandria that so many news articles mourned when Google Books lost is court case.

2.4. Conclusions

I contend that each database is best understood as a negotiation between the noncommercial values of textual reproduction and the commercial environment in which institutions much remain financially solvent. Each database has the goal of making valuable information available. After the 1990s, they are particularly influenced by the utopian ideal that digital reproduction at last made textual reproduction free. Each had to contend, however, with the fact that before a text can be reproduced digitally it must be *created* digitally, and that even if the material costs are entirely eliminated (which, of course, they are not) textual creation continues to have costs in labour. In Paddy Bullard's "Digital Humanities and Electronic Resources in the Long Eighteenth Century," which surveys the research completed and the resources used as of 2013, Bullard is also faced with the task of explaining why multiple services interact so poorly. Bullard, too, observes the core tension between public access vs private profit:

> Viewing the field of eighteenth-century digital humanities as a single prospect, it is the contrast between publicly funded, open-access sites, and privately owned, subscription-access resources that is most striking. Each side of the divide has much to learn from the other. Publicly funded academic projects must acquire the pragmatism and ambitiousness of scale that commercial developers have always shown. Commercial developers must adapt themselves more generously to the principles of scholarly openness and accuracy. They might also imitate the inventiveness of the open sector, its adaptability to the demands raised by different kinds of primary media. Both sides recognize the desirability of making their resources interoperable across the divide, and the business of interconnectivity will preoccupy all kinds of digital humanist in the coming decade. (756)

Bullard is correct to note that there are major disjunctions between databases like the *The British Book Trade Index* or careful online editions like *The Proceedings of the Old Bailey, 1674–1913,* compared to massive archives like ECCO. It seems odd, however, to attribute to ECCO *both* “ambitiousness of scale” *and* “pragmatism” as the lessons for noncommercial projects to imitate, since an ambitious scale is only plausibly pragmatic for a project with the money to sustain itself. Even odder is the idea that commercial developers might voluntarily choose to “adapt themselves more generously to the principles of scholarly openness and accuracy,” when the core business model of a private enterprise relies on its lack of openness, and the private access only seems worth purchasing when its marketers suppress all nuance about accuracy. As Bode observes, “the commercial imperatives of these enterprises arguably depend on them presenting these collections as comprehensive” (Bode World 47). In other words, Bullard has observed an underlying system of profit and non-profit in awkward competition, and examined the outputs of these systems in order to articular their particular virtues and describe what a ‘best of both worlds’ might look like if both parts of the system sought to collaborate together on how best to achieve maximally useful scholarly resources. What Bullard overlooks in this process is that not all parts of this system have the goal of achieving maximally useful scholarly resources.

Bullard suggests tentatively that university presses might be site of bridging efforts between the non-profit and for-profit worlds, but where we can actually see an example occurring is in the Text Creation Partnership. The TCP attempted to intervene in the system with “a public-private partnership, led by libraries” (TCP, “About”); their materials emphasize the “librarian’s attitude toward content” which prioritizes the widest possible access and use. This “librarian’s attitude” is most evident in the (eventual) availability of all of the transcriptions in the public domain, despite the fact that the images they are based on remain privately restricted by the companies which own them. Their description of the “partnership,” however, continues to show signs of the strain in value systems when commercial and noncommercial goals are intertwined: “Through our partnership with private vendors, we had access to a huge trove of images from which to transcribe. In return, these companies were supplied with a full-text index to their images —work which would have otherwise been difficult or expensive to produce.” In other words, through purchasing a service (access to images), the academic institutions received that service. These academic institutions carried out an enormous feat of labour at their own expense, using the service they purchased. Then, “in return,” they provided the results of their labour to the company, for the company to then further profit from the improvements to their service. Most telling, here, is the word “otherwise” in calling this “work which would have otherwise been difficult or expensive to produce.” The suggestion here is that, without the TCP, the companies themselves would not have been willing to undertake the encoding that was so desired by the users of their service. However, the TCP certainly did not make the task any less difficult or expensive. Instead, academic institutions absorbed the difficulty and expense on those companies’ behalf. I do not say that they were wrong to do so: on the contrary, the “librarian’s attitude” mirrors my own attitude, and it is surely to everyone’s benefit for a wonderful thing to exist even if that wonderful thing is not profitable. Rather, I highlight this rhetorical moment in the TCP's self description to suggest that it takes two to collaborate, and that no amount of effort on the librarians' part can change the core institutional drive of a private company. Companies like Gale are perfectly happy to help achieve maximally useful scholarly resources if doing so it also a good way to turn a profit, but this does not mean that they have the same goals as academic institutions. One of the three key aims of the TCP identified on the homepage is to "collaborate with commercial providers, rather than constantly bargaining and competing with them" (TCP "Welcome"). However, the TCP seems instead to have simply come up with a *better* bargain, one which creatively offers scholarly labour as a bargaining chip.

What do these database histories mean for scholars of eighteenth century literature? First and foremost, these histories provide another reminder that these things do not exist prior to interpretation or intervention. It is not merely that they are *shaped* or *influenced* by their institutional contexts, implying small quirks or edge cases which can generally be ignored: they are *constituted in the first place* by those institutional contexts. Secondly, these histories suggest a course of action to be taken in response to the specific institutional factors constituting each database. Scholars periodically acknowledge the gaps between historical events as they occurred and the specific archive, database, or corpus that they are using as a proxy for the idealized concept of "the historical record," but these acknowledgements typically take the form of a statement that some form of bias is assumed to exist, but that this bias is so unknowable and unavoidable that naturally we will just continue onward as if it was not present. Identifying the specific institutional process that led to the current digital infrastructure undermines efforts to brush off these details as unknowable: directly investigating the actual demographics of each resource's holdings, as I do in future chapters, can also render these biases no longer unavoidable.

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.

Arondekar, Anjali. “Without a Trace: Sexuality and the Colonial Archive.” *Journal of the History of Sexuality*, volume 14, number 1/2, Special Issue: Studying the History of Sexuality: Theory, Methods, Praxis, January-April 2005, pp. 10-27. *JSTOR*, http://www.jstor.com/stable/3704707.

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.

Bamman, David, Jacob Eisenstein, and Tyler Schoebeler. “Gender identity and lexical variation in social media.” Journal of Sociolinguistics, volume 18, issue 2, 2014, pp. 135-160. doi:10.1111/josl.12080.

Barthes, Roland. “The Reality Effect.” 1968. *The Rustle of Language*, translated by Richard Howard, Hill and Wang, 1986, pp. 141-148.

Baskin, Jon. “On the Hatred of Literature.” *The Point*, issue 21, 26 January 2020, /web/20200506015431/https://thepointmag.com/letter/on-the-hatred-of-literature/.

Battis, Jes. “Molly Canons: The Role of Slang and Text in the Formation of Queer Eighteenth-Century Culture.” Lumen, volume 36, 2017, pp. 129-141. doi:10.7202/1037858ar.

Bauder, Julia. “HathiTrust as a Data Source for Researching Early Nineteenth-Century Library Collections: Identification, Coverage, and Methods,” *Information Technology and Libraries*, volume 38, issue 4, December 2019. *ProQuest*, ProQuest document ID 2336298791.

Bayard, Pierre. *How to Talk About Books You Haven't Read*, translated by Jeffrey Mehlman. Bloomsbury, 2007.

Behrendt, Stephen C. “Charlotte Smith, Women Poets and the Culture of Celebrity.” *Charlotte Smith in British Romanticism*, edited by Jacqueline Labbe. Pickering & Chatto, 2008, pp. 189-202.

Benatti, Francesca and Justin Tonra. “English Bards and Unknown Reviewers: A Stylometric Analysis of

Thomas Moore and the Christabel Review.” Brea: A Digital Journal of Irish Studies, 7 Oct 2015, web.archive.org/web/20191107181606/https://breac.nd.edu/articles/english-bards-and-unknown-reviewers-a-stylometric-analysis-of-thomas-moore-and-the-christabel-review.

Bernbaum, Ernest. Review of *Charlotte Smith, Poet and Novelist (1749-1806)* by Florence May Anna Hilbish. *Modern Language Notes*, vol. 59, no. 2, 1944, pp. 137–139. *JSTOR*, www.jstor.org/stable/2910610.

Blanch, Anna Maree. *A Reassessment of the Authorship of the* Cheap Repository Tracts. Master’s thesis, Baylor University, 2009.

Blank, Antje. “Charlotte Smith.” Edited by Janet Todd. *The Literary Encyclopedia*, volume 1.2.1.06: *English Writing and Culture of the Romantic Period, 1789-1837*, edited by Daniel Cook and Daniel Robinson, 23 June 2003, www.litencyc.com. Accessed 05 June 2019.

Blaney, Jonathan. “Introduction to the Principles of Linked Open Data.” *The Programming Historian*, volume 6, 2017, web.archive.org/web/20200228212040/https://programminghistorian.org/en/lessons/intro-to-linked-data. Accessed 28 February 2020.

Blayney, Peter W. M. “The Alleged Popularity of Playbooks.” *Shakespeare Quarterly*, vol. 56, no. 1, 2005, pp. 33–50. *JSTOR*, www.jstor.org/stable/3844025.

Blevins, Cameron and Lincoln Mullen. "Jane, John ... Leslie? A Historical Method for Algorithmic Gender Prediction.” *Digital Humanities Quarterly*, volume 9, number 3, 2015, http://www.digitalhumanities.org/dhq/vol/9/3/000223/000223.html

Bode, Katherine. “The Equivalence of ‘Close’ and ‘Distant’ Reading; Or, Toward a New Object for Data-Rich Literary History.” *Modern Language Quarterly*, volume 78, number 1, 2017, pp. 77–106, doi:10.1215/00267929-3699787.

—. *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press, 2018.

Bogost, Ian. *Persuasive Games: The Expressive Power of Videogames*, MIT Press, 2010.

Bourrier, Karen, and Mike Thelwall. “The Social Lives of Books: Reading Victorian Literature on Goodreads.” *The Journal of Cultural Analytics*, 20 February 2020. doi: 10.22148/001c.12049.

Brewer, David. “Counting, Resonance, and Form, A Speculative Manifesto.” Eighteenth-Century Fiction, volume 24, issue 2, 2011, pp. 161-170, doi: 10.1353/ecf.2011.0053.

Brown, Susan, Patricia Clements, Isobel Grundy, Stan Ruecker, Jeffery Antoniuk, and Sharon Balazs. “Published Yet Never Done: The Tension Between Projection and Completion in Digital Humanities Research.” *Digital Humanities Quarterly*, volume 3, number 2, 2009, digitalhumanities.org/dhq/vol/3/2/000040/000040.html

Bruhm, Steven. “The Gothic Novel and the Negotiation of Homophobia.” *The Cambridge History of Gay and Lesbian Literature*, edited by E.L. McCallum and Mikko Tuhkanen, Cambridge UP, 2014, pp. 272-87.

Bullard, Paddy. “Digital Humanities and Electronic Resources in the Long Eighteenth Century.”

Literature Compass, volume 10, 2013, pp. 748-760.

Burton, Matt, Matthew J. Lavin, Jessica Otis, and Scott B. Weingart. “Digits: Two Reports on New Units of Scholarly Publication.” *The Journal of Electronic Publishing*, volume 22, issue 1, 2019, doi:10.3998/3336451.0022.105.

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. “The Dawn of a Philosophy of Visualization.” *Data Visualization in Society*, Amsterdam UP, 2020.

—. “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.

Carretta, Vincent. Unchained Voices: An Anthology of Black Authors in the English-Speaking World of the Eighteenth Century. UP of Kentucky, 2003.

—. “Who Was Francis Williams?” *Early American Literature*, volume 38, number 2, 2003, pp. 213-23. *JSTOR*, jstor.com/stable/25057315.

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.

Chatterjee, Ronjaunee, Alicia Mireles Christoff, and Amy R. Wong. “Undisciplining Victorian Studies.” *LA Review of Books*, 10 July 2020, web.archive.org/web/20200728034955/https://lareviewofbooks.org/article/undisciplining-victorian-studies.

Christy, Matthew, Anshul Gupta, Elizabeth Grumbach, et al. “Mass Digitization of Early Modern Texts With Optical Character Recognition.” *ACM Journal on Computing and Cultural Heritage*, volume 11, number 1, article 6, December 2017, 25 pp., doi:10.1145/3075645.

Chun, Wendy Hui Kyong. “Queerying Homophily.” *Pattern Discrimination*, by Clemens Apprich, Wendy Hui Kyong Chun, Florian Cramer, and Hito Steyerl, meon press and U Minnesota P, 2018, pp. 59-97, doi:10.14619/1457. In Search Of Media series, edited by Götz Bachman, Timon Beyes, Mercedes Bunz, and Wendy Hui Kyong Chun.

Civale, Susan. “Women’s life writing and reputation: A case study of Mary Darby Robinson.” *Romanticism*, vol. 24, no. 2, 2018, pp. 181-202.

Clemens, Justin. “Aggressively middling: The Bourgeois & Distant Reading by Franco Moretti.” Sydney Review of Books, 16 July 2013, web.archive.org/web/20200309032758/https://sydneyreviewofbooks.com/review/aggressively-middling/ Accessed 8 March 2020.

Clery, E.J. The Rise of Supernatural Fiction 1762-1800. Cambridge UP, 1995.

Christman, Paul. “The Cinema of Inadvertence.” *The Hedgehog Review: Critical Reflections on Contemporary Culture*, volume 21, number 3, Fall 2019, web.archive.org/web/20191118012126/https://hedgehogreview.com/issues/eating-and-being/articles/the-cinema-of-inadvertence-or-why-i-like-bad-movies

Cronin, Richard. The Politics of Romantic Poetry: In Search of the Pure Commonwealth. Palgrave Macmillan, 2000.

Cross, Ashley. “From *Lyrical Ballads* to *Lyrical Tales*: Mary Robinson's Reputation and the Problem of Literary Debt.” *Studies in Romanticism*, vol. 40, no. 4, 2001, pp. 571–605, doi:10.2307/25601532.

—. Mary Robinson and the Genesis of Romanticism: Literary Dialogues and Debts, 1784–1821. Routledge, 2016.

Cohen, Margaret. *The Sentimental Education of the Novel*. Princeton UP, 1999.

Coker, Cait, and Kate Ozment. “Building the *Women in Book History Bibliography*, or Digital Enumerative Bibliography as Preservation of Feminist Labor.” *Digital Humanities Quarterly*, volume 13, number 3, 2019, www.digitalhumanities.org/dhq/vol/13/3/000428/000428.html.

Cooke, Richard. “Wikipedia Is the Last Best Place on the Internet.” *Wired*, 17 February 2020, web.archive.org/web/20200227222807/https://www.wired.com/story/wikipedia-online-encyclopedia-best-place-internet.

Cordell, Ryan. “Talking about Viral Texts Failures.” *ryancordell.org*, 25 June 2020, web.archive.org/web/20200625185504/https://ryancordell.org/research/VT-database-fail. Accessed 25 June 2020.

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.

DeLuna, D. N. “Race Matters in the Very Long Eighteenth Century.” *Huntington Library Quarterly*, volume 64, number 3/4, 2001, pp. 507-518, https://www.jstor.org/stable/3817924.

Denham, Alison. “Making Sorrow Sweet: Emotion and Empathy in the Experience of Fiction.” *Affect and Literature*, ed. Alex Houen, Cambridge UP, 2020, pp. 190-210, doi:10.1017/9781108339339.011.

Digital Proofreaders wiki. “General Workflow Diagram.” web.archive.org/web/20200730033032/https://www.pgdp.net/wiki/DP_Official_Documentation:General/General_Workflow_Diagram. Accessed 29 July 2020.

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, https://web.archive.org/web/20190621121928/http://www.textcreationpartnership.org:80/tcp-ecco. Accessed 20 June 2019.

"Eighteenth Century Collections Online (ECCO) FAQ's." Gale Group (?), circa 2008, https://web.archive.org/web/20200603031058/http://find.galegroup.com/ecco/component/researchtools/xml/ecco_35.xml. Accessed 2 June 2020.

Emre, Merve. "Public Thinker: Leah Price on Books, Book Tech, and Book Tattoos." Interview with Leah Price. *Public Books*, 26 Sept 2019, web.archive.org/web/20190928023628/https://www.publicbooks.org/public-thinker-leah-price-on-books-book-tech-and-book-tattoos. Accessed 27 Sept 2019.

Ezell, Margaret J. M. "Big Books, Big Data, and Reading Literary Histories." *Eighteenth-Century Life*, volume 41, issue 3, September 2017, pp. 3-19. *Project MUSE*, muse.jhu.edu/article/667488.

Facer, Ruth. "Ann Radcliffe (1764-1823)." Women Writer Biographies, Chawton House, chawtonhouse.org/the-library/library-collections/womens-writing-in-english/women-writer-biographies/.

Falkovich, Jacob. "100 Ways to Live Better." Put A Number On It!, 30 December 2019, web.archive.org/save/https://putanumonit.com/2019/12/30/100-ways-to-live-better/.

Farmer, Alan B., and Zachary Lesser. "Structures of Popularity in the Early Modern Book Trade." Shakespeare Quarterly, vol. 56, no. 2, 2005, pp. 206–213. JSTOR, www.jstor.org/stable/3844307.

—. "The Popularity of Playbooks Revisited." *Shakespeare Quarterly*, vol. 56, no. 1, 2005, pp. 1–32. *JSTOR*, www.jstor.org/stable/3844024.

Felski, Rita. "Everyday Aesthetics." the minnesota review, 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.

Fleming, Catherine. *Translation, Reputation, and Authorship in Eighteenth-Century Britain*. PhD Dissertation, University of Toronto, 2018, hdl.handle.net/1807/101619.

Flores, Pepe. "Is Wikipedia the largest-ever digital humanities project? Exploring an emerging relationship." Wikimedia Foundation blog, 17 April 2016, web.archive.org/web/20200428033423/https://blog.wikimedia.org/2016/08/17/wikipedia-largest-digital-humanities-project/

Forster, Chris. "A Walk Through the Metadata: Gender in the HathiTrust Dataset." 8 Sept. 2015, cforster.com/2015/09/gender-in-hathitrust-dataset. Accessed 3 Sept. 2019.

Fowers, Alyssa. “Profiling protest data (or, what I did on my summer vacation).” *Data and Dragons*, 10 Sept 2019, dataanddragons.wordpress.com/2019/09/10/profiling-protests-or-what-i-did-on-my-summer-vacation. Accessed 10 Sept 2019.

Frank, Marcie. “Melodrama and the Politics of Literary Form in Elizabeth Inchbald’s Works.” *Eighteenth-Century Fiction*, volume 27, number 3-4, Spring-Summer 2015, pp. 707-730.

Freedgood, Elaine. “Reading Things.” The Ideas in Things: Fugitive Meaning in the Victorian Novel, U Chicago P, 2006.

Frow, John. *Genre*. Routledge, 2015.

Fry, Carrol L. *Charlotte Smith*. Twayne’s English Authors Series, edited by Herbert Sussman, Twayne, 1996.

Gadd, Ian. “The Use and Misuse of *Early English Books Online*.” *Literature Compass*, volume 6, issue 3, 2009, pp. 680-692, doi:10.1111/j.1741-4113.2009.00632.x.

Gale. “Eighteenth Century Collections Online.” web.archive.org/web/20200324195501/https://www.gale.com/primary-sources/eighteenth-century-collections-online Accessed 24 March, 2020.

Gamer, Michael. Romanticism and the Gothic: Genre, Reception, and Canon Formation. Cambridge UP, 2000.

Gamer, Michael, and Terry F. Robinson. “Mary Robinson and the Dramatic Art of the Comeback.” *Studies in Romanticism*, vol. 48, no. 2, 2009, pp. 219–56. *JSTOR*, www.jstor.org/stable/25602191.

Garnai, Amy. Revolutionary Imaginings in the 1790s: Charlotte Smith, Mary Robinson, Elizabeth Inchbald. Palgrave Macmillan, 2009.

Garrett, Jeffrey. “Subject Headings in Full-Text Environments: The ECCO Experiment.” *College & Research Libraries*, volume 68, number 1, 2007, doi: 10.5860/crl.68.1.69.

Garside, Peter. “The English Novel in the Romantic Era: Consolidation and Dispersal.” *The English Novel 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles*, edited by Peter Garside, James Raven, and Rainer Schöwerling, vol. 2: 1800-1829, edited by Peter Garside and Rainer Schöwerling with Christopher Skelton-Foord and Karin Wünsche. Oxford UP, 2000.

Garside, Peter, James Raven, and Rainer Schöwerling, editors. The English Novel 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles, Oxford UP, 2000. 2 vols.

—. General Introduction. *The English Novel 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles*, edited by Peter Garside, James Raven, and Rainer Schöwerling, Oxford UP, 2000. 2 vols.

Gavin, Michael. “Historical Text Networks: The Sociology of Early English Criticism.” *Eighteenth-Century Studies*, volume 50, number 1, 2016, pp. 53-80.

Gilbert, Geoff. “The Durability of Affect and the Ageing of Gay Male Queer Theory.” *Affect and*

Literature, ed. Alex Houen, Cambridge UP, 2020, pp. 133-158, doi: 10.1017/9781108339339.008.

Goldie, Mark and Robert Wokler, editors. *The Cambridge history of eighteenth-century political thought*. Cambridge UP, 2006.

Gonda, Caroline. "Review of *Heteronormativity in Eighteenth-Century Literature and Culture*, ed. by Ana de Freitas Boe and Abby Coykendall." *Eighteenth Century Studies*, volume 49, issue 3, 2016, pp. 427-428.

Google Books. "Google Books History." Site as of 25 March 2020, web.archive.org/web/20200326031915/https://books.google.com/googlebooks/about/history.html Accessed 25 March 2020.

—. "Google Books History." Site as of 6 February 2016, https://web.archive.org/web/20160206043510/http://books.google.com/googlebooks/about/history.html. Accessed 25 March 2020.

—. "Google Books Ngram Viewer: Datasets." 2013, https://web.archive.org/web/20200608004518/http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

—. "Google Books Ngram Viewer: Info." 2013, https://web.archive.org/web/20200608004424/https://books.google.com/ngrams/info.

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.

Gray, Kishonna L. Review of *Algorithms of Oppression: How Search Engines Reinforce Racism*. *Feminist Media Studies*, volume 19, issue 2, pp. 308-310, doi:10.1080/14680777.2019.1579984.

Gregg, Stephen H. "Finding ECCO-TCP texts." *Manicule: Thoughts on the Eighteenth Century, Daniel Defoe, and Digital Humanities*, Wordpress, 16 Aug. 2017, shgregg.com/2017/08/16/finding-ecco-tcp-texts. Accessed 20 June 2019.

Guillory, John. Cultural Capital: The Problem of Literary Canon Formation. U Chicago P, 1993.

Hammond, Adam. Literature in the Digital Age: An Introduction. Cambridge UP, 2016.

Hane, Paula. "Project Gutenberg Progresses." *Information Today*, volume 21, number 5, May 2004, web.archive.org/web/20200325195437/http://www.infotoday.com/it/may04/hane1.shtml. Accessed 25 March 2020.

Hart, Michael. "The History and Philosophy of Project Gutenberg." *Project Gutenberg*, August 1992. web.archive.org/web/20200312224522/https://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart Accessed 12 March 2020.

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

—. “Governance.” *HathiTrust*, web.archive.org/web/20200325212223/https://www.hathitrust.org/governance. Accessed 25 March 2020.

—. “Help - Copyright.” *HathiTrust*, web.archive.org/web/20200325212528/https://www.hathitrust.org/help_copyright%23RestrictedAccess. Accessed 25 March 2020.

—. “Major Library Partners Launch HathiTrust Shared Digital Repository.” *HathiTrust*, 13 October 2008, web.archive.org/web/20200325213628/https://www.hathitrust.org/press_10-13-2008. Accessed 25 March 2020.

—. “Member Community.” *HathiTrust*, web.archive.org/web/20190915204844/https://www.hathitrust.org/community. Accessed 15 Sept 2019.

—. “Our Digital Library.” *HathiTrust*, web.archive.org/web/20190915204611/https://www.hathitrust.org/digital_library. Accessed 15 Sept 2019.

—. “Our Membership.” *HathiTrust*, web.archive.org/web/20190915204720/https://www.hathitrust.org/partnership. Accessed 15 Sept 2019.

Heisel, Andrew. “Hannah More’s Art of Reduction.” *Eighteenth-Century Fiction*, volume 25, number 3, Spring 2013, pp. 557-588.

Hosch, William L. “Project Gutenberg.” *Encyclopedia Britannica*, 17 August 2017, web.archive.org/web/20200325190246/https://www.britannica.com/topic/Project-Gutenberg. Accessed 25 March 2020.

Hunt, Bishop C. “Wordsworth and Charlotte Smith.” *The Wordsworth Circle*, vol. 1, no. 3, 1970, pp. 85. *ProQuest*, ProQuest Document ID 1300171026.

IPUMS NAPP. “What is NAPP?” *North Atlantic Population Project*, web.archive.org/web/20200209014205/https://www.nappdata.org/napp/intro.shtml. Accessed February 8, 2020.

Jockers, Matthew. Macroanalysis: Digital Methods and Literary History. U Illinois P, 2013.

Juxta Commons. “A User Guide to Juxta Commons.” web.archive.org/web/20200227014953/http://juxtacommons.org/guide.

Karels, Liene. “HathiTrust adds new members, goes global.” *Montage: Arts + Creativity*, University of Michigan, November 2010, web.archive.org/web/20140302084528/http://www.montage.umich.edu/2010/11/hathitrust-adds-new-members-goes-global. Accessed 25 March 2020.

Karian, Stephen. "The Limitations and Possibilities of the ESTC." *The Age of Johnson*, vol. 21, 2011, pp. 283-297. *ProQuest,* ProQuest document ID 1689625001.

King, Kathryn R. “Introduction: Hans Turley, Queer Studies, and the Open-Hatched Eighteenth Century.” *The Eighteenth Century*, volume 53, number 3, 2012, pp. 265-272. *JSTOR*, www.jstor.org/stable/23365012.

Klein, Lauren. “Distant Reading After Moretti.” *Arcade: Literature, the Humanities, & the World*, 2018, web.archive.org/save/https://arcade.stanford.edu/blogs/distant-reading-after-moretti. Accessed 19 September 2019.

Klein, Ula, and Emily MN Kugler. “Eighteenth-Century Camp Introduction.” *ABO: Interactive Journal for Women in the Arts, 1640-1830*, volume 9, issue 1, 2019, pp. 1-12. doi:10.5038/2157-7129.9.1.1180

Korshin, Paul J. Review of Bibliography, Machine Readable Cataloguing, and the ESTC. A Summary History of the Eighteenth Century Short Title Catalogue. Working Methods. Cataloguing Rules. A Catalogue of the Works of Alexander Pope Printed Between 1711 and 1800 in the British Library, by R. C. Alston and M. C. Jannetta. Eighteenth-Century Studies, volume 12, number 2, 1978, pp. 209–212. JSTOR, www.jstor.org/stable/2738046.

Labbe, Jacqueline. “Introduction.” *Charlotte Smith in British Romanticism*, edited by Jacqueline Labbe. Pickering & Chatto, 2008, pp. 1-11.

—. “Selling One's Sorrows: Charlotte Smith, Mary Robinson, and the Marketing of Poetry.” *The Wordsworth Circle*, volume 25, number 2, 1994, pp. 68-71. *ProQuest*, ProQuest Document ID 1300173031.

LaGuardia, Cheryl. Review of *Eighteenth Century Collections Online*. *Library Journal*, May 2004, pp. 123-124.

Lahti, Leo, Niko Ilomäki, and Mikko Tolonen. “A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800.” *LIBER Quarterly*, volume 25, number 2, 2015, pp. 87–31, doi:10.18352/lq.10112.

Lebert, Marie. “Project Gutenberg (1971-2008).” *Project Gutenberg*, May 2008, www.gutenberg.org/cache/epub/27045/pg27045-images.html Accessed 12 March 2020

Lee, Haimin. “15 Years of Google Books.” *The Keyword*, Google.

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.

Lincoln, Matthew D. "Tidy Data for the Humanities." *Matthew Lincoln, PhD* (blog), 26 May 2020, https://web.archive.org/web/20200529025634/https://matthewlincoln.net/2020/05/26/tidy-data-for-%20humanities.html

Liu, Alan. “Where is Cultural Criticism in the Digital Humanities?” *Debates in the Digital Humanities*. Edited by Matthew K. Gold. University of Minnesota Press, 2012.

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.

Manshel, Alexander, Laura B. McGrath, and J.D. Porter. “Who Cares About Literary Prizes?” Public Books, 3 August 2019, www.publicbooks.org/who-cares-about-literary-prizes

“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/.

McInnes, Andrew. “Should We Cancel Romantic Studies?” *The Romantic Ridiculous*, 15 June 2020, web.archive.org/web/20200627220515/https://romanticridiculous.wordpress.com/2020/06/15/should-we-cancel-romantic-studies.

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.

Michel, Jean-Baptiste et al. “Quantitative Analysis of Culture Using Millions of Digitized Books.” *Science*, volume 331, issue 6014, 2011), pp. 176-182. doi:10.1126/science.1199644.

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.

—. “Network Theory, Plot Analysis.” *New Left Review*, volume 68, 2011, pp. 80–102.

Mullen, Lincoln. “gender: Predict Gender from Names Using Historical Data.” *R* package version 0.5.2. *GitHub*, https://github.com/ropensci/gender

Mullen, Lincoln, Cameron Blevins, and Ben Schmidt. “Package ‘gender.’” November 9, 2019.

Murphie, Andrew. “The Digital’s Amodal Affect.” *Affect and Literature*, ed. Alex Houen, Cambridge UP, 2020, pp. 390-407, doi: 10.1017/9781108339339.022.

Murphy, Peter. Poetry as an occupation and an art in Britain, 1760-1830. Cambridge UP, 1993.

Nicolazzo, Sarah. “Reading Clarissa’s “Conditional Liking”: A Queer Philology.” Modern Philology, volume 112, 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.

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.

OpenRefine. Version 3.1, Nov. 29, 2018, openrefine.org.

Otchenash, Maiia. “Suing Online Platforms for Copyright Infringements: The Choice of Court and Law in the ‘Project Gutenberg’ Scenario.” *IDP Revista de Internet, Derecho y Politica*, no. 29, 2019, doi:10.7238/idp.v0i29.3175.

O’Quinn, Daniel. “Half-History, or The Function of Cato at the Present Time.” *Georgian Theatre in an Information Age: Media, Performance, Sociability*, special issue of *Eighteenth-Century Fiction*, vol. 27, no. 3-4, 2015, pp. 479-507. *Project Muse*, muse.jhu.edu/article/584623.

Potter, Franz J. The History of Gothic Publishing, 1800-1835: Exhuming the Trade. Palgrave Macmillan, 2005.

Price, Leah. The Anthology and the Rise of the Novel: From Richardson to George Eliot. Cambridge UP, 2000, doi:10.1017/CBO9780511484445.

Prior, Karen Swallow. “Hannah More.” *The Literary Encyclopedia*, volume 1.2.1.06: *English Writing and Culture of the Romantic Period, 1789-1837*, edited by Daniel Cook and Daniel Robinson, 16 Dec. 2008, www.litencyc.com. Accessed 29 June 2019.

Project Gutenberg. “Credits.” 7 June 2006, web.archive.org/web/20200325192439/https://www.gutenberg.org/wiki/Gutenberg:Credits. Accessed 25 March 2020.

—. “The CD and DVD Project.” 19 August 2017, web.archive.org/web/20200325194042/https://www.gutenberg.org/wiki/Gutenberg:The_CD_and_DVD_Project. Accessed 25 March 2020.

—. “Feeds.” 1 April 2020, https://web.archive.org/web/20200608003736/http://www.gutenberg.org/wiki/Gutenberg:Feeds.

—. “Offline Catalogs.” 6 January 2020,

https://web.archive.org/web/20200608003535/http://www.gutenberg.org/wiki/Gutenberg:Offline_Catalogs.

—. "Partners, Affiliates and Resources." 24 January 2019, web.archive.org/web/20200325193013/https://www.gutenberg.org/wiki/Gutenberg:Partners,_Affiliates_and_Resources. Accessed 25 March 2020.

Punter, David. Review of Cultural Capital: The Problem of Literary Canon Formation, by John Guillory. Non-Standard Englishes and the New Media, special issue of The Yearbook of English Studies, volume 25, 1995, pp. 229-230. JSTOR, www.jstor.org/stable/3508832.

Raven, James. "Historical Introduction: The Novel Comes of Age." *The English Novel 1770-1829: A Bibliographical Survey of Prose Fiction Published in the British Isles,* edited by Peter Garside, James Raven, and Rainer Schöwerling, vol. 1: 1770-1799, edited by James Raven and Antonia Forster with Steven Bending, Oxford UP, 2000.

Readings, Bill. Review of *Cultural Capital: The Problem of Literary Canon Formation*, by John Guillory. *Modern Language Quarterly,* volume 55, number 3, Sept 1994, pp. 321-326, doi:10.1215/00267929-55-3-321.

@RealSaavedra (Ryan Saavedra). "Socialist Rep. Alexandria Ocasio-Cortez (D-NY) claims that algorithms, which are driven by math, are racist." *Twitter*, 22 Jan. 2019, 3:27 AM, web.archive.org/save/https://twitter.com/realsaavedra/status/1087627739861897216?lang=en.

Reason. "How to Politely Download All English Language Text Format Files from Project Gutenberg." *Ex Ratione*, 1 November 2014, web/20200506014409/https://www.exratione.com/2014/11/how-to-politely-download-all-english-language-text-format-files-from-project-gutenberg/

Reinert, Thomas. Review of *Cultural Capital: The Problem of Literary Canon Formation*, by John Guillory. *Modern Fiction Studies*, volume 42, number 1, 1996, pp. 221-224. *ProQuest*, ProQuest Document ID 2152910014.

Richardson, Leonard. "Project Gutenberg Books Are Real." *Books in Browsers V Proceedings,* volume 18, issue 1, winter 2015, doi:10.3998/3336451.0018.126.

Riddell, Allen, Troy J. Bassett, Laura Schneider, et al. "Common Library 1.0: A Corpus of Victorian Novels Reflecting the Population in Terms of Publication Year and Author Gender." arXiv:1909.02602v1 [cs.DL]

Rigby, Mair. "Uncanny recognition: Queer theory's debt to the Gothic." Gothic Studies, vol. 11, no. 1, 2009, pp. 46-57.

Roberts, Bethan. "Charlotte Smith: Elegiac Sonnets, and Other Essays." Edited by Daniel Robinson. *The Literary Encyclopedia,* volume 1.2.1.06: *English Writing and Culture of the Romantic Period, 1789-1837*, edited by Daniel Cook and Daniel Robinson, 02 January 2014, www.litencyc.com. Accessed 05 June 2019.

Rogers, Deborah D. Ann Radcliffe: A Bio-Bibliography. Bio-Bibliographies in World Literature, number 4. Greenwood Press, 1996.

Rosenberg, Scott. “How Google Book Search Got Lost.” *Wired*, 11 April 2017, web.archive.org/web/20200326035223/https://www.wired.com/2017/04/how-google-book-search-got-lost/. Accessed 25 March 2020.

Runge, Laura. “Mary Darby Robinson (1758?-1800) - Bibliography.” chuma.cas.usf.edu/~runge/MRobinson.htm. Accessed 07 June 2019.

Sandvoss, Cornel. “The Death of the Reader: Literary Theory and the Study of Texts in Popular Culture.” *Fandom: Identities and Communities in a Mediated World*, edited by Jonathan Gray, C. Sandvoss, and C. Lee Harrington. NYU Press.

Seaver, Nick. “Bastard Algebra.” *Data, Now Bigger and Better!*, edited by Tom Boellstorff and Bill Maurer. Prickly Paradigm, 2015.

Sedgwick, Eve 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.

Sentance, Nathan. “Your neutral is not our neutral.” *Archival Decolonist*, 27 September 2019, web.archive.org/web/20200621053700/https://archivaldecolonist.com/2019/09/27/your-neutral-is-not-our-neutral-3/

Shaw, Zed. *Learn Python the Hard Way*, Addison-Wesley Professional, 2013.

Shirky, Clay. “Why Abundance is Good: A Reply to Nick Carr.” *Encyclopædia Britannica Blog*, 17 July 2008, blogs.britannica.com/2008/07/why-abundance-is-good-a-reply-to-nick-carr. Accessed 10 Sept 2019.

Shiroma, Yumi Dineen. “Conjectures on World Literature, Revisited.” June 3, 2018, web.archive.org/web/20200309032430/http://yumidineenshiroma.org/blog/conjectures-on-world-literature-revisited/ Accessed March 1, 2020.

Sinanan, Kerry. “Heterogeneous Blackness: Peter Brathwaite’s Eighteenth-Century Re-portraits.” *The 18th-Century Common*, 13 July 2020, web.archive.org/web/20200721052125/https://www.18thcenturycommon.org/blackportraiture. Accessed 18 July 2020.

Somers, James. “Torching the Modern-Day Library of Alexandria.” *The Atlantic*, 20 April 2017, web.archive.org/web/20200326035404/https://www.theatlantic.com/technology/archive/2017/04/the-tragedy-of-google-books/523320/. Accessed 25 March 2020.

Spedding, Patrick. “‘The New Machine’: Discovering the Limits of ECCO.” *Eighteenth-Century Studies*, volume 44, issue 4, 2011, pp. 437-453.

St. Clair, William. *The Reading Nation in the Romantic Period*. Cambridge UP, 2007.

Stanton, Judith Phillips. “Recovering Charlotte Smith’s Letters: A History, With Lessons.” *Charlotte Smith in British Romanticism*, edited by Jacqueline Labbe. Pickering & Chatto, 2008, pp. 159-173.

Stott, Anne. “Hannah More Chronology.” *The Victorian Web*, 2002, www.victorianweb.org/authors/more/chron.html.

Suarez, Michael F. “Towards a bibliometric analysis of the surviving record, 1701–1800.” *The Cambridge History of the Book in Britain, Volume 5: 1695–1830*, edited by Michael F.

Suarez and Michael L. Turner, Cambridge UP, 2009, pp. 39-65.

Syme, Holger Schott. “Imaginary Targets.” *Los Angeles Review of Books*, 5 Nov 2012. web.archive.org/web/20191023050529/https://lareviewofbooks.org/article/in-defense-of-data-responses-to-stephen-marches-literature-is-not-data.

Tabor, Stephen. "ESTC and the Bibliographical Community." *The Library: The Transactions of the Bibliographical Society*, vol. 8 no. 4, 2007, pp. 367-386. *Project MUSE*, muse.jhu.edu/article/230381. (If I add new Tabor citations, go back and clarify which ones I’m already quoting)

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

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

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

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

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

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

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

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

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

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

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

—. “Our scholarly partners.”

web.archive.org/web/20200311042125/https://textcreationpartnership.org/about-the-tcp/about-partner-libraries/. Accessed 10 March 2020.

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

Thaventhiran, Helen. “Feelings under the Microscope: New Critical Affect.” Affect and Literature, ed. Alex Houen, Cambridge UP, 2020, pp. 83-99, doi:10.1017/9781108339339.005.

Townshend, Dale and Angela Wright. “Gothic and Romantic engagements: The critical reception of Ann Radcliffe, 1789–1850.” *Ann Radcliffe, Romanticism and the Gothic*, Cambridge University Press, 2014.

Tufte, Edward. 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.

—. Distant Horizons: Digital Evidence and Literary Change. U Chicago P, 2019.

—. “Do humanists get their ideas from anything at all?” *The Stone and the Shell*, 24 Jan 2012, web.archive.org/web/20191105004437/https://tedunderwood.com/2012/01/24/discovery-and-hypothesis-testing. Accessed 4 Nov 2019.

—. Why Literary Periods Mattered: Historical Contrast and the Prestige of Literary Studies. Stanford UP, 2013.

Underwood, Ted and David Bamman. “The instability of gender.” The Stone and the Shell, 9 January 2016, https://web.archive.org/web/20200207044340/https://tedunderwood.com/2016/01/09/the-instability-of-gender/. Accessed 6 Feb 2020.

—. “Preregistered Hypotheses for Evaluating Models of Literary Character,” June 2014, hdl.handle.net/2142/49936.

University Libraries, University of Colorado Boulder. “Increasing Access with Google Books.” 13 November 2019, /web/20200608004707/https://www.colorado.edu/libraries/2019/11/13/increasing-access-google-books.

van Hooland, Seth, Ruben Verborgh, and Max De Wilde. “Cleaning Data with OpenRefine.” *The Programming Historian*, edited by Adam Crymble, 5 August 2013, web.archive.org/web/20200719213948/https://programminghistorian.org/en/lessons/cleaning-data-with-openrefine. Accessed 19 July 2020.

Vander Meulen, David. “*ESTC* as Foundational and Always Developing.” *The Age of Johnson*, vol. 21, 2001, pp. 263-282.

Vara, Vauhini. “Project Gutenberg Fears No Google.” *The Wall Street Journal Online*, 10 December

2005, https://www.wsj.com/articles/SB113415403113218620.

Vareschi, Mark and Mattie Burkert. “Archives, Numbers, Meaning: The Eighteenth-Century Playbill at Scale.” *Theatre Journal*, volume 68, issue 4, 2016, pp. 597-613. *ProQuest*, ProQuest Document ID 1886308168.

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, https://web.archive.org/web/20200605030011/https://mattwilkens.com/2015/01/13/literary-attention-lag/. Accessed 5 June 2020.

Yeager, Myron D. “Articulating Male Homosexual Identity in the Long Eighteenth Century.” Review of The Overflowing of Friendship: Love between Men and the Creation of the American Republic by Richard Godbeer, The Gendering of Men, 1600–1750: Volume 2, Queer Articulations by Thomas A. King, and Perverse Romanticism: Aesthetics and Sexuality in Britain, 1750–1832 by Richard C. Sha. ANQ: A Quarterly Journal of Short Articles, Notes, and Reviews, volume 23, number 4, pp. 259–267, 2010. doi:10.1080/0895769X.2010.517094.

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.