Abstract: To what extent do the material historical conditions of digitization lead to unrepresentative samples in digital archives of literary works? Examining records from the English Short Title Catalogue, Eighteenth Century Collections Online, HathiTrust, and the Text Creation Partnership for works published in England between 1789 and 1799, I do not replicate findings in other periods that more selective databases underrepresent works by women. Instead, it is the unsigned author who fails to garner institutional support, revealing the extent to which literary studies remains dominated by the idea of the author as an individual.

# History of the Book Database:

# How Four Digital Archives of Eighteenth-Century WorksSelect for Gender

Despite the impression they give of always being ‘up to date’, digital resources for literary study often reflect much earlier conceptions of literature. It has now been more than thirty years since feminist criticism started having major successes in eighteenth century studies. No eighteenth centuryist is now likely to say that Charlotte Smith, for example, is irrelevant or unimportant to the period, a stance supported by her substantial presence in traditional scholarly infrastructures. Smith 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, and conference panels, and every other sign that she is a major figure. But what about digital infrastructures? Eighteenth Century Collections Online (ECCO), for example, began in 2003, well after figures like Charlotte Smith had been ‘recovered,’ but the microfiche images underlying ECCO date from the 1980s. ECCO might easily reflect 1980s priorities about what kinds of works are worth investing in. To what extent do research infrastructures, especially digital ones, lag behind scholarly consensus? Beyond just one author, to what extent do uneven digitization practices shape scholarly access to eighteenth century texts?

Scholarly concerns regarding what is lost in a shift to digital books often focus on physical features of the books themselves, such as the texture of the paper or the relative size of a duodecimo versus a folio. Another form of absence, however, is the book which has not been digitized. The precise scope and nature of these absences are difficult to define except in very small samples. A recent experiment by Allen Riddell and Troy J. Bassett, for example, uses exhaustive bibliographies of new British novels in 1836 and 1838, to determine what percentage of these works today have digital editions available.[[1]](#footnote-2) They find that novels by women, and single-volume novels, are substantially less likely to have been digitized than multi-volume novels by men in these two years. Narrowing their focus by genre, year, and geographic region makes it possible to have a reasonably complete initial list. They conclude cautiously that “it seems prudent for researchers to assume that gathering samples of fiction or non-fiction works published in the British Isles during the 1830s and 1840s from the major digital libraries will yield samples which do not reflect the population.”[[2]](#footnote-3) It is by combining multiple partial experiments of this kind that we will eventually be able to describe, and appropriately respond to, uneven digitization on a larger scale. Examining a different sliver, works published in England between 1789 and 1799, unexpectedly, I have *not* found the same bias against the digitization of women’s writing for works — though I do not therefore conclude that digital archives do accurately reflect the original population of published works.[[3]](#footnote-4) Riddell and Bassett search Internet Archive, HathiTrust, Google Books, and the British Library, and consider a novel to be “available digitally” if it is present in any one of these repositories, but these resources are not interchangeable. I contend that it is also crucial to distinguish between *which* digital archives contain different works. Doing so highlights the fact that any database is shaped by the material conditions of its creation, and that uneven digitization is not as straightforward as a simple gender bias.

To more fully map the terrain of uneven digitization in eighteenth century studies, I begin with the English Short Title Catalogue (ESTC) as the largest and most comprehensive set of records: for works printed in England 1789-99, they list 51,090 titles.[[4]](#footnote-5) Eighteenth Century Collections Online, for the same period, has facsimiles of 26,848 works— barely more than half the titles listed in the ESTC. HathiTrust, a different collection of facsimiles, has 8,220 works. The smallest and most specialized academic text repository, the Text Creation Partnership (TCP) collection of meticulously hand-encoded transcriptions, contains only 525 works from the 1790s: an almost negligible 1% of the ESTC. The shrinking size of each resource is a natural result of the different level of investment required to add a work to each. The ESTC might be assumed to be the least influenced by selection bias, since it includes the least information about each work and has the largest commitment to comprehensiveness. Correspondingly, the TCP is likely the most influenced by selection bias, with the highest per-book cost and the fewest claims to represent a complete whole. Comparing the two raises the question: which kinds of works were selected to receive scholarly investment? If scholars adopt TCP texts for computational distant-reading, how distorted is their sample compared to the original population of works published in the eighteenth century? Looking beyond what might seem merely technological or methodological: what do the resources we have created reveal about our discipline? In posing these questions, rather than questions that might be more directly literary, I bring the emerging field of digital infrastructure studies to the eighteenth century, to undertake a history of the book database.

## Charlotte Smith in the Digital Archive

To better understand the current state of eighteenth-century works in digital databases, it is helpful to follow a single figure and the specific reasons that their works have been included or excluded over the years, beginning with the pre-digital creation and social valuation of those works. Charlotte Smith is now famous in multiple areas of eighteenth-century literature: as an early Romantic poet, for her genre-bending *Elegiac Sonnets* which influenced Wordsworth and for her longer works *The Emigrants* and *Beachy Head*; as a Gothic and political novelist, for works like *The Old Manor House* and the radical *Desmond*; and as a pioneer in children’s educational literature, for her natural history. During her lifetime, she achieved critical success as a poet and commercial success as a novelist, though her radical politics jeopardized her ability to sell novels later in her career. Smith’s posthumous reception has undergone multiple reversals of fortune. Shortly after her death in 1806, Smith was widely eulogized and anthologized, remembered and emulated as an important British poet.[[5]](#footnote-6) As the nineteenth century went on, poetesses began to be anthologized separately from poets, in collections with ambitions that were commercial rather than prestigious; Smith, too, “lost intellectual ground” even as she continued to be sold.[[6]](#footnote-7) By the end of the nineteenth century, even these volumes marginalized Smith’s poetry, which was dismissed as trite and depressing.

In the early twentieth century, Florence Hilbish produced the first extensive study of Smith, considering her as both poet and novelist, in 1941,[[7]](#footnote-8) to unappreciative reviews: Ernest Bernbaum’s faint praise said that “much time and care have been devoted to it; whether deservedly, is perhaps questionable.”[[8]](#footnote-9) Louise Duckling credits the feminist movement of the 1960s and 1970s with the beginning of Smith’s recovery: the renewed interest in women’s writing rediscovered her novels, and especially the radical political content which Hilbish had observed.[[9]](#footnote-10) 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; he called for further study of Smith, but Smith’s importance remained a reflection of Wordsworth’s importance rather than borne of the value of her own verses.[[10]](#footnote-11) Through the 1980s, Smith was often treated tokenistically as an interesting woman novelist, and a minor pre-Romantic poet, before rising to independent prominence in the 1990s. With work by Stephen Curran, Roger Lonsdale, Jennifer Breen, Andrew Ashfield, Jacqueline Labbe, and many others, Smith entered the 2000s as central not only to the feminist canon of women writers, but to Romanticism itself.

How has this varying reception history impacted Smith’s presence in contemporary databases? Our baseline list of “all works published by Smith,” against which we can compare each resource to locate its absences, includes 47 editions published in her lifetime or in the year immediately following her death.[[11]](#footnote-12) None of the four databases I consider (ESTC, ECCO, HathiTrust, or TCP) include all 47. Figure 1 shows smaller and smaller databases winnowing down her full output arbitrarily:

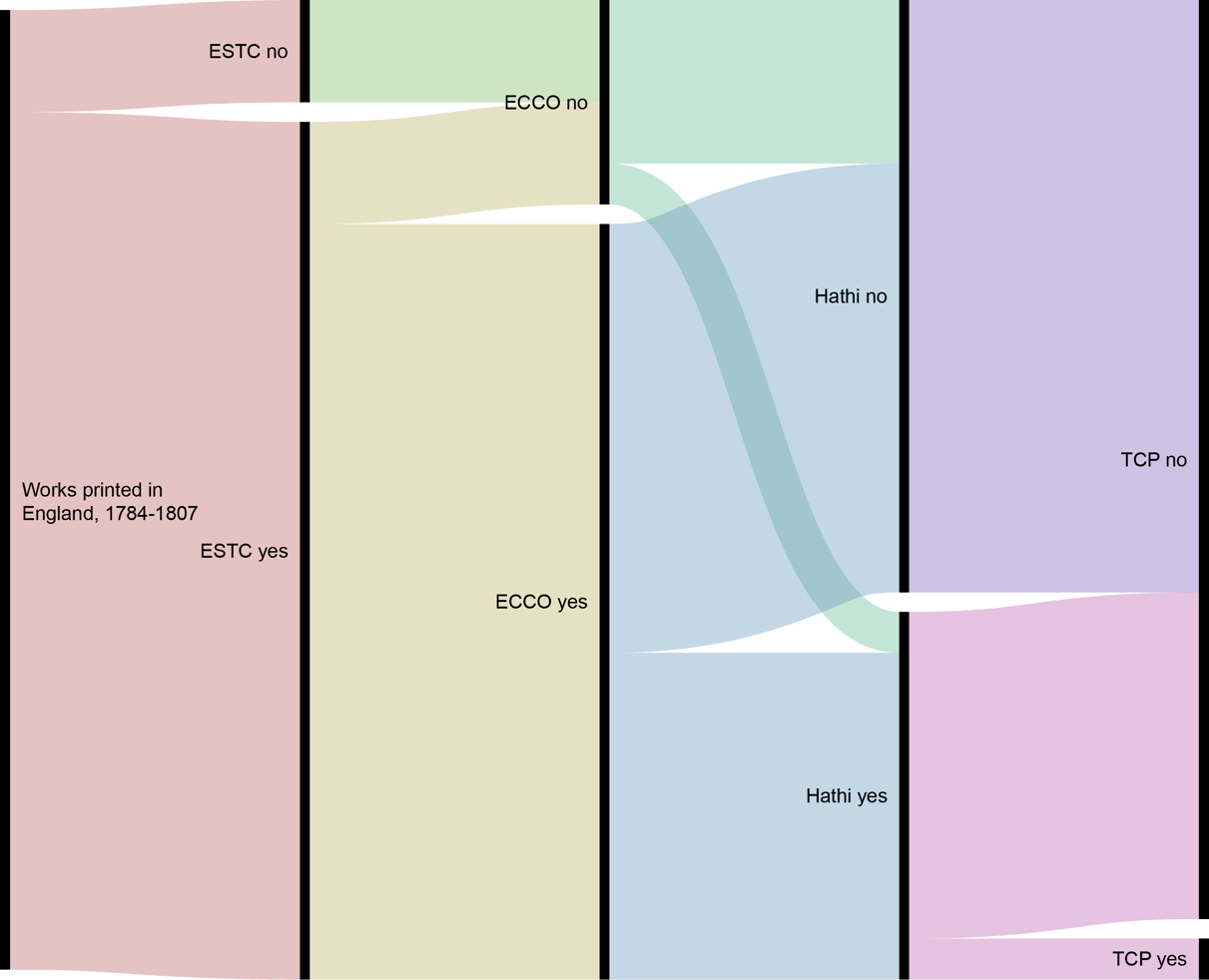


Figure 1: An alluvial chart, showing the winnowing down of Smith’s works from database to database.

What, then, is not selected for preservation? For both ESTC and ECCO, the answer begins with a selection criteria at the site of collection: neither includes any works published after 1800. The ESTC, therefore, 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 then adds a selection bias at the site of digitization. Five works which fall within ECCO’s collection mandate are nonetheless excluded: 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). Presumably, these volumes were too scarce, too inaccessible, or too fragile to be imaged by Research Publications, Inc when they undertook to produce the *Eighteenth Century Collection* microfilm in 1983.

HathiTrust’s images come not from 1980s microfilms, but from Google Books scans undertaken from 2004 to today. This scanning was undertaken without reference to another bibliography, or any selection criteria over than availability: Google has periodically announced an intention to scan every book in the world. In order to induce university libraries to allow Google to borrow books by the semi-truck, Google provided universities with their own copies of any scans produced from that library’s collection; HathiTrust brings together those libraries’ copies into an academic resource. Although the same jpegs are involved in both Google Books and HathiTrust, therefore, they remain separate institutions, with Google Books still decidedly corporate while HathiTrust is a non-profit academic institution. 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).[[12]](#footnote-13) 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.[[13]](#footnote-14) 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.

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. Looking more closely at these two works, as the only works by Charlotte Smith which are available in all four databases, illuminates how none of the four truly offer access to the “same” work. First, of course, there is the obvious detail that they provide fundamentally different information. ESTC gives bibliographic records and library shelf-marks, deferring outward to a ‘real’ physical book. The other three choose a specific copy to represent this ‘real’ book, and provide a mediated reformulation of that copy. ECCO provides monochrome scans of microfiche images; HathiTrust, full-colour digital photographs of book pages. The TCP offers meticulously hand-encoded transcripts. 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 copy used by the others), which is 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. 60p. ; 4⁰” with the same note “numbers 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. However, the data on page numbers is now misleading, as HathiTrust omits the detail about the volume’s inaccurate pagination.[[14]](#footnote-15) It becomes clear that HathiTrust, borne as it is from Google’s corporate priorities, has lost effort on reinventing the wheel: even as HathiTrust fills some gaps in existing scholarly resources, it does so at the expense of having to start from scratch, creating new gaps where previous work makes them needless.

In the end, the lacunae of the digital resources have led to Smith’s presence being smaller than might be expected — even in resources created more recently, after her ‘recovery’ might have encouraged more investment in the preservation and recirculation of her works. However, Smith’s absences are not easily attributed to systemic sexism, or even to a bias against ‘unserious’ genres which can often function as a veiled sexism. Instead, the missing works tell us much more about the databases that lack them than they tell us about scholarly embrace of Smith herself. Examining the history of those databases in greater detail, we can see how their current collections have been shaped by technological change, and by the need for the organizations behind each collection to compete with each other in particular moments.

## Databases of Eighteenth Century Literature, 1977-2020

These four databases — ESTC, ECCO, HathiTrust, and TCP — are only part of the larger ecosystem of digital book repositories. These databases require description preliminary to analysis, since they have conventionally been treated as tools for accessing objects of study, rather than objects of study in themselves. Here, I will give a brief overview of how the holdings of these resources relate to each other, and then a detailed chronological history of their development from the 1970s to the present. Discussing the specifics of digitization recontextualizes these resources as historical artifacts. Figure 2 shows a Venn diagram of the approximate relative scale, and overlap in holdings, of eight databases in three ecosystems. 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 more commercial Google-backed ecosystem: Google Books, Google Ngrams, and HathiTrust.[[15]](#footnote-16) Project Gutenberg is also pictured for comparison, to indicate the presence of a third ecosystem of textual collection, not-for-profit but also not-for-academics; its complex case is beyond the scope of this article, but offers an important consideration for future work.

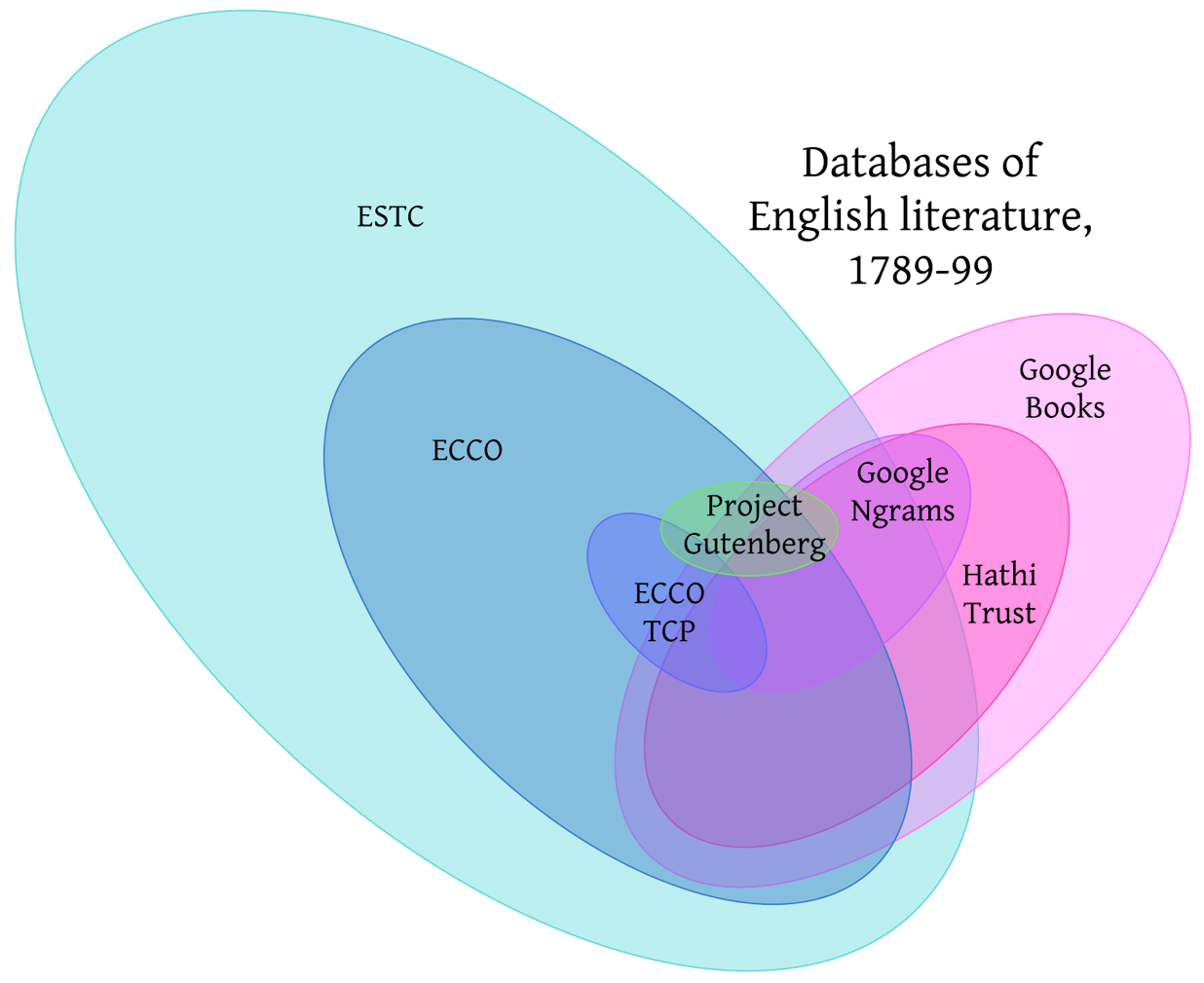


Figure 2: A hand-drawn Venn diagram of the general relative scale and overlapping holdings of databases containing 1790s literature.

Databases within a shared ecosystem may or may not be interoperable, but they made their initial textual selections with a similar logic. An immediate difference between them, for example, is that the Google ecosystem treats textual holdings like a trade secret; Google Books and Google Ngrams as impossible to discuss except through the academically-inflected proxy of HathiTrust. 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. A chronological organization contrasts with essentially teleological descriptions of individual resources, which tend to work backward from a current state to present a 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.

Important threads of this history of mass digital archives begin long before the invention of the computer, with the bibliographic collections which served as the extant solutions to the challenge of large-scale text-tracking. *The Pollard and Redgrave Short-Title Catalogue for 1476–1640* first appeared in 1926. Donald Wing’s catalogue for 1641–1700 appeared in 1951. After the completion of Wing’s STC, exploratory studies throughout the 1950s and 60s pursued the feasibility of accounting for the much larger body of printed work produced in the eighteenth century.[[16]](#footnote-17) Concurrently, Wing’s seventeenth-century STC underwent redevelopment into a second edition, which published its first volume in 1976.[[17]](#footnote-18) These print bibliographies establish the expectations for large-scale resources: some concrete scope will be defined geographically and chronologically, and the resource will be consistent and comprehensive within that scope. Specialized vocabularies will capture details about the works (such as their size or pagination), and all included works will be evaluated according to the same standards. Perfection is understood to be infeasible, but comprehensiveness and consistency remain the core principles. In 1977, these principles entered the digital realm with the Eighteenth Century Short Title Catalogue, which shared the print catalogues’ ambitious simplicity: to accurately describe every edition of every printed work in English or from the United Kingdom in the eighteenth century. Unlike earlier catalogues, the Eighteenth-Century Short Title Catalogue was conceived as a digital database from the beginning. In 1980, with cataloguing barely begun, the ESTC began to go online.[[18]](#footnote-19) The possibility of disseminating updates digitally created the environment for a hybrid kind of short-title catalog, which attempted a bibliography’s consistency without (yet) its comprehensiveness.

Meanwhile, microfilm remained a central archival technology. The book facsimiles which would become Eighteenth Century Collections Online (ECCO) began in 1983, when the company Research Publications, Inc began to produce its *Eighteenth Century Collection* microfilm. Research Publications was founded in 1981 as a for-profit company. 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; over the course of ECCO’s history, the private company that owns the microfilms changes frequently. Concurrently with these new microfilms, the ESTC began to make its contents available in non-digital form. 1983 saw the first stable publication of the ESTC, a fiche catalogue and index of the British Library’s holdings.[[19]](#footnote-20) The next microform publication from the ESTC, a joint Anglo-American interim publication of the ESTC file which was expected in 1984, did not appear.[[20]](#footnote-21) 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 proofreading and in-person examination of books, which increased the labour of verifying records. Although he consistently warned “how easily strategic decisions based exclusively on cost usually lead to greater, not less, eventual costs,” the ESTC each year faced a new budget struggle.[[21]](#footnote-22) This created a paradox: 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. It also highlights the defining importance of funding to academic resources.

It is in light of these funding difficulties that we might interpret the ESTC’s approach to online access in 1985, when the online databases in RLIN and BLAISE were upgraded to allow dynamic updates and continuous access to a single shared file.[[22]](#footnote-23) To facilitate its use, the ESTC distributed a manual for searching the online file, but although the manual was free, use of the ESTC itself was not. Institutions or individuals paid to subscribe to the ESTC itself, paid per query for searches to be run, paid per minute for connecting to the database, and often paid for access to the computers they used in their 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,” framing these new forms of cost in the context of already normalized academic infrastructure.[[23]](#footnote-24) Despite his suggestion of old ways being superseded by new “cyber-equivalents,” however, these printed “predecessors” continued as active contemporaries of the ESTC database, with the second volume of the second edition of the STC was published in 1986. Despite its struggles, the ESTC’s ambitions grew, and in 1987 it expanded its scope to also include all printed books prior to the eighteenth century, changing its name to the English Short Title Catalogue.

The 1990s, which saw an explosion of activity for crowdsourced group-knowledge projects like Project Gutenberg and Wikipedia, were quieter for eighteenth-century academic databases, despite the popularization of the internet. The second edition of the print STC completed its publication in 1991, which marked the end of the Bibliographical Society’s ability of to sponsor it.[[24]](#footnote-25) In 1999, the ESTC took over responsibility for hosting and updating STC data.[[25]](#footnote-26) The late 90s also saw the beginning of Early English Books Online (EEBO), when, in 1998, ProQuest (formerly University Microfilms) began to make digitised copies of its microfilms available through the internet. The next year, the Text Creation Partnership was formed, a collaboration between university libraries and ProQuest to produce accurate machine-readable transcripts of EEBO texts. ECCO emerged four years after EEBO, when, in 2003, Thomson Gale (formerly Research Publications, Inc) made digital copies of *Eighteenth Century Collection* microfilms available online. This timing suggests that when one company introduces a new resource in order to gain a competitive edge (as EEBO did in reselling its book images in a new format), it sets a new norm for its competitors. Just as ECCO followed EEBO in offering page images online, so too did they follow EEBO in attempting manual transcription of their collections. In 2005, the TCP began encoding ECCO texts, as an experimental alternative to transcripts produced by Optical Character Recognition (OCR) technologies. In both cases, the timing suggests that eighteenth-century texts grow more accessible to scholars as a response to developments in EEBO, and Gale’s commercial competition with ProQuest.

In 2004, Google Print was announced, and for the first time in thirty years, new scans of books were produced at scale. Google negotiated with libraries to access what they estimated at more than 15 million volumes, and set the goal to scan them all. Eighteenth century works were included almost incidentally. In 2006, the ESTC was made available to search for free online, and began adding full title and imprint transcripts to its listing. Meanwhile, Google had already scanned millions of new books. In 2008, however, Google begin to face legal repercussions for their “scan first and ask questions later” approach to mass digitization.[[26]](#footnote-27) At first these legal challenges did not slow things down— Google Books began the 2010s by announcing, in mid-2010, that they had scanned more than 12 million books and intended to scan all known existing 129,864,880 books within a decade. At the end of 2010, they launched the Google Ngram viewer, as a research tool borne of all their scanning. Meanwhile, however, their legal court case continued, and scanning began to slow in 2012. In April 2016, the US Supreme Court ruled in Google’s favor. Despite their victory— or perhaps, because they had already achieved their true goals, which were unrelated to making books accessible— Google essentially ended their investment in Google Books. They no longer brag about the number of works scanned, and certainly do not circulate their intention of scanning every book ever published. Much has been written outside academic circles about the odd half-death of Google Books, and more than one journalist has taken the position that academics and librarians are responsible for killing what could have been an enormous public good.[[27]](#footnote-28) In the long term, its impact in scholarly circles has been to create the conditions for HathiTrust, by leaving individual libraries as custodians of photographs of their own books. These libraries have a clearly valuable resource in the images, but Google did not create the images with libraries in mind. As was evident in our case study of Charlotte Smith above, Google often duplicated scholarly work by re-digitizing books that already existed in facsimile. More problematically, Google hindered the ability of book historical research to use its facsimiles by maintaining very little information about them. Google Books itself is infamous for inaccurate dates;[[28]](#footnote-29) HathiTrust matches books to library catalogue records to provide more information than Google’s baseline, but there is a true missed opportunity in the lack of clear links between the Google ecosystem of digitized books and the academic ecosystem.

The outcome of these individual companies’ attempts, over the last fifty years, to turn a profit or secure funding through documenting eighteenth-century works, is the ecosystem we began with in Figure 2. In that figure, I defined each digital repository based on the overlap in the actual eighteenth-century works each one included. The TCP was a subset of ECCO, which was itself a subset to the ESTC; and, partially overlapping with those works, HathiTrust’s collection was a subset of the Google Books collection. It is also possible, however, to organize these repositories according to their commercial status, as in Figure 3. In many ways, this is a much less functional figure. It is no longer possible, for example, to conceive of overlaps between the texts within each resource. Instead, what we discover is that these digital resources no collection — neither the the ESTC-derived holdings nor the Google-derived images — competes with itself in a financial niche. Project Gutenberg no longer sticks out as an outlier; instead, it is the TCP which forges an unusual alternative path as both commercial and non-commercial.[[29]](#footnote-30) This figure also suggests an explanation for why it is so difficult to identify which works appear in both Google Books and ECCO: closer integration would muddy their commercial strategies, which sell distinct products to distinct customer bases.[[30]](#footnote-31) In contrast, it now seems less potentially-surprising that HathiTrust has an ongoing project to connect its holdings to ESTC records, since non-commercial entities compete (if they compete at all) on different terms. Granting organizations can be swayed by claims to groundbreaking originality, but they can also be swayed by proven utility; the ESTC and HathiTrust need not fight a zero-sum game.

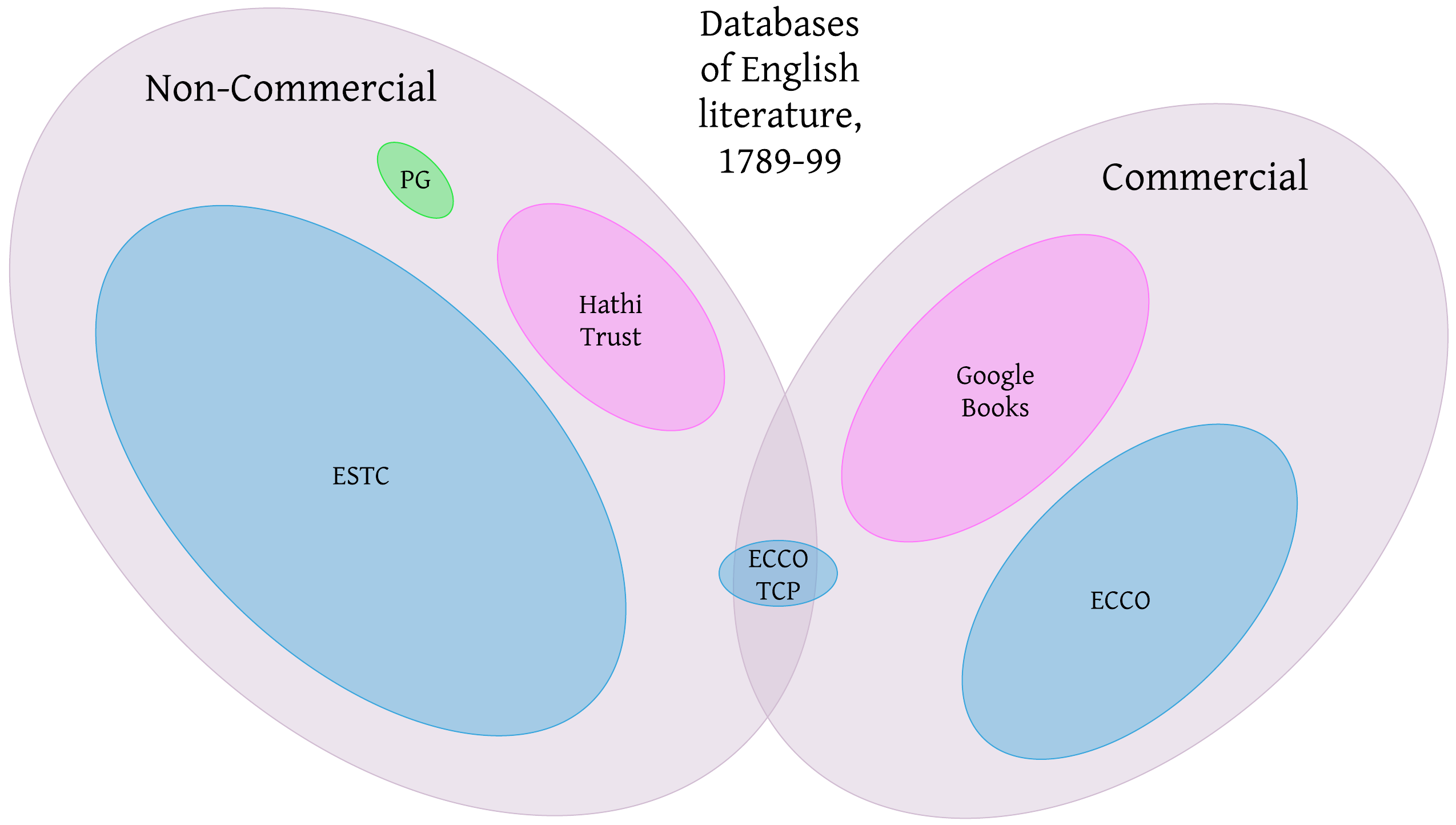


Figure 3: A hand-drawn Venn diagram of the general relative scale of databases containing 1790s literature, organized into the categories of “commercial” and “non-commercial.” No overlaps between their actual holdings are depicted, since doing so occludes the distinction between commercial and non-commercial. Holdings derived from the ESTC are still in blue, and holdings derived from Google in pink, with the crowdsourced Project Gutenberg in green.

## Databases of Eighteenth-Century Literature, 1789-99

Returning to the eighteenth century, and how it has been digitized: the baseline hypothesis, investigating a historical phenomenon, is that factors like systemic sexism are likely to have an effect. As I have shown above, the digital archives currently used by eighteenth-century scholars were created during their own historic moments, which influenced their priorities as texts were digitized. Resources which provide more information about particular works — such as ECCO, which provides PDF scans of microfiche, or the TCP, which provides detailed XML-encoded transcripts — inevitably provide their more thorough information about fewer texts. The ESTC is able to include so many works in its database in part because it includes so little about each work: not much more than the information found on the title page, and a list of libraries where the original can be consulted. Google Books (the source of HathiTrust’s images) achieved its scale in part by dramatically reducing the detail and precision of each work’s records. As the more resource-intensive archives created their digital surrogates, therefore, the texts they include are likely to be less indiscriminate, and more strongly influenced by a personal assessment of what works are most important. In many other areas, writing by women has been dismissed as less important or worthy of serious study than writing by men. If we compare the holdings of major digital archives, do the smaller and more resource-intensive archives show evidence of similar choices to systematically opt out of an investment in women’s writing?

Guillory’s *Cultural Capital* proposes that, if female authors, for example, do not make up a demographically accurate 50% of our literary syllabi, this is not necessarily because women’s writing has been excluded or repressed.[[31]](#footnote-32) Guillory draws a distinction between rhetoric “construing the process of canon formation as an exclusionary process essentially the same as the exclusion of socially defined minorities from power”[[32]](#footnote-33) — that is, the patriarchal suppression of works by women, which can be addressed by now ‘representing’ the excluded works within the canon — versus what he sees as the real cause, a historical fact that “women were routinely excluded from *access to literacy*, or were proscribed from composition or publication in the genres considered to be serious rather than ephemeral.”[[33]](#footnote-34) Applied to the digital database, the first explanation, which might be termed the ‘exclusion hypothesis,’ would manifest in the form of massive archives with large numbers of works by men and women, which get filtered through exclusionary scholarly processes into smaller, selective repositories where men now make up an oversized share. In the second explanation, which might be termed the ‘absence hypothesis,’ works by women will have a diminished role from the very beginning, since women were prevented from writing in the first place. To see if I can identify one of these two phenomena in eighteenth-century databases, I have acquired records for all works published in England[[34]](#footnote-35) between 1789 and 1799[[35]](#footnote-36) in the ESTC,[[36]](#footnote-37) ECCO,[[37]](#footnote-38) HathiTrust,[[38]](#footnote-39) and the TCP.[[39]](#footnote-40) After removing misclassified works which were from outside England or outside the 1790s, the ESTC contained 51,090 titles, ECCO 26,848 titles, HathiTrust 8,220 titles, and the TCP only 525 titles. ECCO, in all its capaciousness, therefore contains only 52.5% of the titles listed in the ESTC. The TCP provides access to an almost negligible 1% of the works published. Figure 4 shows the relative author demographics of each database.[[40]](#footnote-41)

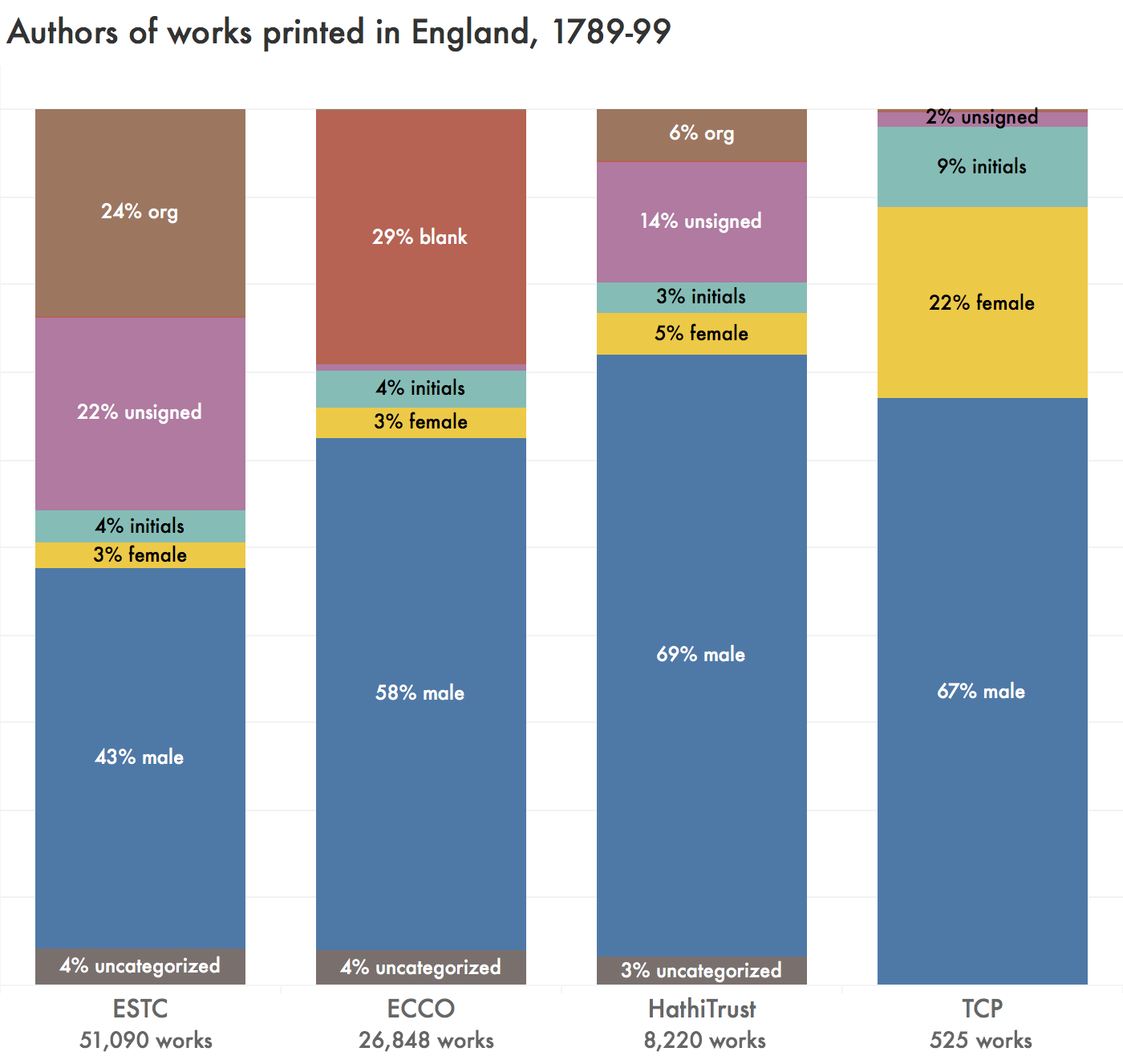


Figure 4: The relative authorship status of works published in England 1789-99, as identified by the ESTC, ECCO, HathiTrust, and the TCP. Works by “org” are attributed to groups or organizations as their authors. “Unsigned” works have no information in the author field. “Blank” in ECCO captures both unsigned works and works by organizations.

Smaller, less comprehensive archives do generally dedicate a higher percentage of their space to male authors. 43% of the 51,090 titles in the ESTC from this decade are identifiably by men,[[41]](#footnote-42) which rises substantially to 69% for HathiTrust and 67% for the TCP. However, the increase in male authors is not at the expense of identifiable female authors, who make up a stable or increasing percentage of the works as databases grow more specialized— including an astonishing 22% of the TCP corpus. The initial number for women, of course, is extremely low: a mere 3%. This is far below the percentage of women found in genre-specific surveys. For example, during the years 1789-99, 20.2% of the new novels written in England were attributed to female names on their title pages and prefaces.[[42]](#footnote-43) The ESTC and ECCO have many medical texts, sermons, natural histories, legal opinions, and other genres from which women are *de facto* excluded. HathiTrust’s slight increase in female authorship might show a collections-level bias where university libraries prefer more recognizably “literary” materials like novels, poetry, and drama, all of which allowed for female authorship. But even if the TCP were assumed to contain exclusively novels, at 22% female authorship it would be statistically overrepresenting women.

Unexpectedly, perhaps, the overlooked group, unable to compete with male authors for archival attention, is not “female authors” but “authors which are not associated with an individual personality.”[[43]](#footnote-44) There are two ways to explain the disappearance of unsigned authors from smaller corpora. The first explanation is Guillory’s ‘exclusion hypothesis,’ that unsigned authors have, essentially, faced discrimination due to their status as unsigned: out of a fixed pool of unsigned works, fewer were chosen for further scholarly dissemination than from the pool of signed works. The second explanation is not, however, the ‘absence hypothesis,’ since the “unsigned author” is not a meaningful real-world identity category; instead, I propose that unsigned authors disappear due to in infrastructural phenomenon that might be called the ‘intervention hypothesis.’ Namely, that the dissemination of a work, or scholarly investment in a work, reduces the likelihood that it will *remain* unsigned in the database. For an author to appear as “unsigned,” they need not only be absent from the title page, but they must also avoid later scholarly attribution of authorship. And indeed, names are placed into the relevant database fields wherever possible.[[44]](#footnote-45) Ann Radcliffe’s first novel, for example, was published unsigned, and the next several editions identified her as the “authoress” of the previous works, and only with the third edition of *The Romance of the Forest* did the name “Ann Radcliffe” appear on the book. All of the earlier editions, however, are consistently linked to Ann Radcliffe as the author. There is obvious practical use in disseminating author identifications — indeed, there is very little point in determining the authors of pseudonymous works if this information is not fully incorporated into the scholarly record. But this particular implementation also occludes how eighteenth century readers actually encountered author information. An identified name overwrites the information recorded on the title page. In this way, scholars infrastructurally eliminate the “unsigned” author.

Rather than finding bias against authors based on the social category of gender, then, I have found a much stronger bias in favour of a particular concept of the author itself. These resources assume, in their very infrastructure, that the “author” must be reducible to a single, known individual. ECCO so thoroughly rejects the idea that organizations could be considered as “authors” that it does not record them in the “author” field. The underrepresentation of unsigned or corporate authors cannot be mapped onto the multicultural rhetoric of political representation, the way Guillory has described with gendered representation.[[45]](#footnote-46) Instead, the meaning of this finding stems from the very *lack* of a broader social importance to “unsigned” or “non-individual” authors as a class of humans: the conclusions we can draw here are not about the world at large, but about the institutional processes of academic research.

## “Representativeness” in Databases

Returning to the idea of “representativeness” — what it means for a database to “represent” or be “representative of” a population — returns us, also, to the still-fraught relationship between distant reading and the conception of the literary canon. Twenty years after Guillory, we are still debating the need for “literary criticism … to conceptualize a new disciplinary domain,”[[46]](#footnote-47) now in the context of computation. The reconceptualization of literary study itself is at the core of Franco Moretti’s coinage of ‘distant reading’: the problem for which “[r]eading ‘more’ seems hardly to be the solution” is the problem of conceiving of a “world” literature, rather than the “canonical fraction, which is not even one per cent of published literature.”[[47]](#footnote-48) His 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. Katherine Bode, for example, in “The Equivalence of ‘Close’ and ‘Distant’ Reading,” argues that Franco Moretti and Matthew Jockers[[48]](#footnote-49) replicate the approaches of New Criticism with their corpora, and calls for “a new scholarly object of analysis”[[49]](#footnote-50) that directly examines historical and textual context of corpora as representations of “literary systems.”[[50]](#footnote-51) Lauren 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.”[[51]](#footnote-52) This framing re-creates, at the site of the corpus, the 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, on the surface of it, is not without grounds. Uneven technological accessibility of texts has 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. However, what we learn from these case studies, of Charlotte Smith and of the 1790s at large, is that the decisions made about inclusion and exclusion from digital databases simply do not mirror the decisions made in syllabi or anthologies. 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 must 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,” 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.[[52]](#footnote-53)

Bullard is correct that there are major disjunctions between careful online editions like *The Proceedings of the Old Bailey, 1674–1913* and 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 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 private access only seems worth purchasing when its marketers suppress nuance about accuracy. As Bode observes, “the commercial imperatives of these enterprises arguably depend on them presenting these collections as comprehensive.”[[53]](#footnote-54) In other words, Bullard has observed an underlying system of profit and non-profit organizations in awkward competition, examined the outputs of each approach in order to articulate their particular virtues, and finally described 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 that university presses might bridge the non-profit and for-profit worlds, but we can also see an example in the Text Creation Partnership. The TCP intervened in the system with “a public-private partnership, led by libraries;” their materials emphasize the “librarian’s attitude toward content” which prioritizes the widest possible access and use.[[54]](#footnote-55) This “librarian’s attitude” is most evident in the (eventual) availability of all transcriptions in the public domain, despite the fact that the images transcribed 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.[[55]](#footnote-56) 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 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 is that, without the TCP, the companies would not have been willing to undertake the encoding desired by their customers. However, the TCP certainly did not make the task any less difficult or expensive.[[56]](#footnote-57) 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 is also a good way to turn a profit, but this does not mean that they have the same institutional goals as libraries.[[57]](#footnote-58) 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.”[[58]](#footnote-59) 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 scholarly materials 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. Second, 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 can also render these biases no longer unavoidable.

## Conclusion

As archivists and librarians often note, “the archive” is not only a concept to be brought into play with literary theory, but a specific and materially constructed space, shaped by archivist labor and constraints. A digital archive is no different. To explain the current contents of online databases like ESTC, ECCO, HathiTrust, and the TCP, we must look to histories that begin before the internet itself. Recognizing that they are historically constructed resources, we can be alert to selection factors that cause all of them, even the most comprehensive, to differ from the imagined prior whole of “literature” which we might wish for them to represent. More importantly, attention to their particular details reveals that the systemic selection factors that have defined their gaps and exclusions do not always map onto the demographics we dedicate the most attention to recovering. Corporate business models are likely to play as large a role as social biases. Scholarly infrastructure can create new demographics like “the unsigned author” which do not constitute an identity category in the real world. And sometimes success creates its own new puzzles: it is a valuable accomplishment of the TCP, for example, that it has increased the availability of writing by women — but it seems prudent for researchers to assume that using the TCP to gather samples of works published in England during the late eighteenth century will yield samples which do not reflect the population. Unless researchers takes explicit steps to the contrary, these samples will not reflect eighteenth century history, but their own.

1. Riddell, Alan, and Troy J. Bassett, “What Library Digitization Leaves Out: Predicting the Availability of Digital Surrogates of English Novels,” arXiv:2009.00513 [cs.DL], 1 Sep 2020. [↑](#footnote-ref-2)
2. Riddell, Alan, and Troy J. Bassett, “What Library Digitization Leaves Out,” 14. [↑](#footnote-ref-3)
3. As the fuller explanation of my methodology makes clear, my inquiry is broader than Riddell and Bassett’s in both genre and chronology, as I examine all titles and not just novels, and eleven years rather than two. Instead, my concession to feasibility is narrowing my geographical focus from the British Isles to only England. I also examine all publications, not only new publications, which is ironically simpler even though it involves more titles because it spares me the bibliographic task of identifying first editions. [↑](#footnote-ref-4)
4. I treat the ESTC as, if not exactly a complete list of the whole of “things that were published,” at least a *substantial* list against which other resources can be compared. The ESTC intentionally excludes some categories of print production, such as playbills, and might easily exclude other works unintentionally. Accordingly, I make limited claims about how databases differ from eighteenth-century print production itself, focusing on the narrower scope of how they differ from each other. [↑](#footnote-ref-5)
5. 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. [↑](#footnote-ref-6)
6. Duckling, “‘Tell My Name,’” 216. [↑](#footnote-ref-7)
7. Hilbish, Florence May Anna, *Charlotte Smith, Poet and Novelist (1749-1806)*, Ph.D. Dissertation, University of Pennsylvania, 1941. [↑](#footnote-ref-8)
8. Bernbaum, Ernest, review of *Charlotte Smith, Poet and Novelist (1749-1806)* by Florence May Anna Hilbish, *Modern Language Notes*, volume 59, number 2, 1944,138. [↑](#footnote-ref-9)
9. Duckling, “‘Tell My Name,’” 216. [↑](#footnote-ref-10)
10. Hunt, Bishop C. “Wordsworth and Charlotte Smith.” *The Wordsworth Circle*, volume 1, number 3, 1970. [↑](#footnote-ref-11)
11. I include the first year after her death to capture publications which Smith herself prepared for publication, such as *Beachy Head*, but no editions past 1807 in order to focus this experiment on Smith in her own time. [↑](#footnote-ref-12)
12. Volumes 4 and 5 of *Letters of a Solitary Wanderer* are in fact part of the same HathiTrust bibliographic record as the first three volumes. The publication date for the combined five-volume work is listed as “1800-1802.” [↑](#footnote-ref-13)
13. Several of HathiTrust’s records provide “mixed copies” like this, with some volumes scanned from one library’s collection 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 libraries (with no indication of which library provided it). [↑](#footnote-ref-14)
14. Consulting the HathiTrust facsimile shows that it, too, omits the page numbers 9-16, going directly from page 8 to page 17 without a break in the poem. HathiTrust also omits information on the three unnumbered pages between the preface and the poem. Evidently, a human did consult the book, to identify a nine-page preface in roman numerals and the page number on the last page, but they did not carry out a full collation. [↑](#footnote-ref-15)
15. HathiTrust is not a commercial resource in itself, but is depicted as part of the commercial ecosystem because its page images are directly provided by Google’s commercial scanning initiatives. [↑](#footnote-ref-16)
16. Korshin, Paul J. Review of *Bibliography, Machine Readable Cataloguing, and the ESTC*, by R. C. Alston and M. C. Jannetta. *Eighteenth-Century Studies*, volume 12, number 2, 1978, 209. [↑](#footnote-ref-17)
17. Vander Meulen, David. “ESTC as Foundational and Always Developing.” *The Age of Johnson*, volume 21, 2001, 268. [↑](#footnote-ref-18)
18. Vander Meulen, “ESTC as Foundational and Always Developing,” 270. [↑](#footnote-ref-19)
19. Crump, “Short Title Catalogue On-Line,” 105. [↑](#footnote-ref-20)
20. Korshin, review of *Bibliography, Machine Readable Cataloguing, and the ESTC,* 212. [↑](#footnote-ref-21)
21. Alston, Robin. “The Eighteenth Century Short Title Catalogue: A Personal History to 1989,” www.r-alston.co.uk/estc.htm. [↑](#footnote-ref-22)
22. Crump, “Short Title Catalogue On-Line,” 106. [↑](#footnote-ref-23)
23. Tabor, Stephen. “ESTC and the Bibliographical Community.” *The Library: The Transactions of the Bibliographical Society*, volume 8, number 4, 2007, 367. [↑](#footnote-ref-24)
24. Vander Meulen, “ESTC as Foundational and Always Developing,” 269. [↑](#footnote-ref-25)
25. Vander Meulen, “ESTC as Foundational and Always Developing,” 270. [↑](#footnote-ref-26)
26. Google Books’ eight-year legal battle was a strange and complex case — Google very nearly settled out-of-court in an elaborate arbitration which would have effectively redefined copyright law for “orphan works” (copyright-protected work for which rightsholders are positively indeterminate or uncontactable) and given Google a de facto monopoly to sell these works. [↑](#footnote-ref-27)
27. See particularly Somers, “Torching the Modern-Day Library of Alexandria,” *The Atlantic*, 20 April 2017. [↑](#footnote-ref-28)
28. Infamously, many books list “1899” as their publication date because this number was used as a placeholder for “no date.” For more on Google Books’ bibliographical unreliability, see, for example, Harper, Sarah Fletcher, Google Books review, *Journal of Electronic Resources in Medical Libraries*, volume 13, 2016, pp. 2–7; Jacsó, Péter, “Google Scholar revisited,” *Online Information Review*, volume 32, 2008, pp. 102–14; and Weiss, Andrew, “Examining massive digital libraries (MDLs) and their impact on reference services,” *Reference Librarian*, volume 57, 2016, pp. 286–306. [↑](#footnote-ref-29)
29. HathiTrust’s Google-based origins might suggest that it, too, could be conceived of as a commercial-noncommercial hybrid, but HathiTrust as an organization is made up solely of academic institutions. Google itself is not interested in HathiTrust, whereas both ProQuest and Gale are intimately connected with the TCP. [↑](#footnote-ref-30)
30. The strangeness of considering Google Books and ECCO as “competitors” is itself telling: they are the most similar pair of resources in scale, data provided (i.e., page images), and financial model — and yet they seem to occupy entirely different worlds. I suggest here that their strangeness from each other may be conceived of as itself a market strategy. [↑](#footnote-ref-31)
31. Guillory, John, *Cultural Capital: The Problem of Literary Canon Formation,* U Chicago P, 1993. [↑](#footnote-ref-32)
32. Guillory, *Cultural Capital*, 8. [↑](#footnote-ref-33)
33. Guillory, *Cultural Capital,* 15. [↑](#footnote-ref-34)
34. Many place names, such as “Plymouth” or “Halifax,” can refer to locations inside or outside of England. I opted to avoid false negatives, at the expense of including more false positives: when I could not *rule out* England, I assumed the location referred to England. I do not expect false positives to impact my findings: relatively few works had ambiguous publishers, and I have no reason to believe the effects I observe would be different outside of England. [↑](#footnote-ref-35)
35. Many works are assigned ambiguous dates, or ranges of dates. In order to exclude non-1790s works via a consistent rule, I assumed works were published in the earliest year proposed. I would therefore remove a work identified as “1780 to 1790,” but keep works identified as “1791 to 1794” or “1798 to 1800.” [↑](#footnote-ref-36)
36. ESTC data was generously provided by the ESTC Editorial Team in early 2017, and reflects the contents of the ESTC Database as hosted by the British Library at that time. [↑](#footnote-ref-37)
37. MARC records of ECCO holdings were generously provided by the University of Toronto Libraries. [↑](#footnote-ref-38)
38. HathiTrust data is publicly available. I created a HathiTrust “collection” of works published in England 1789-99 and downloaded the related metadata. [↑](#footnote-ref-39)
39. TCP data is publicly available. I am very grateful for the assistance of Brian Gauch in writing a program to extract bibliographic metadata from the XML headers of TCP texts. [↑](#footnote-ref-40)
40. After discovering the limits of automatic methods, I assigned genders manually. Most texts were associated with a small number of common names which presented no interpretive difficulties (John, William, Thomas, James), but as I reached the long tail of publications I researched all necessary individuals to confirm that, for example, Tench Coxe was male and Benedikte Naubert was female. The assumption that first names themselves are “male” or “female” is not one I endorse for all historical periods (it is clearly untenable in our present moment), but in my close encounter with 1790s names, it was well-matched by eighteenth century social practices. [↑](#footnote-ref-41)
41. The category of works “identifiably by men” includes named authors with male first names (e.g., John), semi-unsigned authors whose author attribution clearly signalled male authorship (e.g., “Mr. R”), and pseudonyms which similarly must apply to men (e.g., “a country clergyman”). It excludes authors identified only by their initials, even when a subject expert could match those initials to a specific individual of known gender (e.g., W.H. Ireland). [↑](#footnote-ref-42)
42. I have calculated this figure from the table presented in 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*, vol. 1: 1770-1799, edited by James Raven and Antonia Forster with Steven Bending, Oxford UP, 2000, 47. [↑](#footnote-ref-43)
43. The category “unsigned” captures works to which no name was signed, as well as pseudonymous works which offered no clues to an author’s gender (e.g. “A friend of peace”). “Organizations” include various branches of the government, reports from scholarly and charitable organizations, catalogues from various companies, and other group publications. The identifier “blank,” for ECCO, appears to capture both unsigned works and works by organizations. Works by “initials” include those for which the database has provided an expansion of the initials (e.g., a listing of “W[illiam] H[enry] Ireland”), to reflect the fact that the original author name was of ambiguous gender. An author will be identified as “male” or “female” rather than as “initials” if, in addition to the initials, a gendered title is provided, e.g., “Mrs. R.” The “uncategorized” works are those which I have not manually assessed. [↑](#footnote-ref-44)
44. “Real” names are often provided even when the substitution contains little useful information. One work, for example, titled “Observations on the origin and effects of the Test Act,” has its author listed in the ESTC as “Hudson, active 1790.” A note suggests that this attribution is shaky: “Attribution from Halkett & Laing which seems to be based on a MS attribution on O copy title page, no longer legible.” This is the only item in the ESTC attributed to “Hudson, 1790” so identification does not add value by linking the work to another. In contrast, the title page identifies the author as “a dissenter” — an informative identification which might form the basis of a valuable database query. [↑](#footnote-ref-45)
45. “Unsigned” or “pseudonymous” people are not an identity category in the real world requiring equitable social standing. These authors, if identified, would all turn out to have names, genders, races, class standings, and all the other traits of humans in the world. Their status as “unsigned” does not mean that those traits are absent — it means they are undetermined at the site of the title page. [↑](#footnote-ref-46)
46. Guillory, *Cultural Capital*, 265. [↑](#footnote-ref-47)
47. Moretti, Franco, “Conjectures on World Literature.” *New Left Review*, volume 1, issue 1, 2000, 55. [↑](#footnote-ref-48)
48. Jockers is best known for his monograph, *Macroanalysis: Digital Methods and Literary History*, University of Illinois Press, 2013. [↑](#footnote-ref-49)
49. 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, 79. [↑](#footnote-ref-50)
50. Bode, “Equivalence,” 97. [↑](#footnote-ref-51)
51. Klein, Lauren. “Distant Reading After Moretti,” *Arcade: Literature, the Humanities, & the World*, 2018, arcade.stanford.edu/blogs/distant-reading-after-moretti. [↑](#footnote-ref-52)
52. Bullard, Paddy, “Digital Humanities and Electronic Resources in the Long Eighteenth Century,” *Literature Compass*, volume 10, 2013, 756. [↑](#footnote-ref-53)
53. Bode, Katherine, *A World of Fiction: Digital Collections and the Future of Literary History,* University of Michigan Press, 2018, 47. [↑](#footnote-ref-54)
54. Text Creation Partnership (TCP), “About the partnership,” textcreationpartnership.org/about-the-tcp. Accessed 12 March 2020. [↑](#footnote-ref-55)
55. I have not yet ascertained whether participating TCP institutions were able to purchase access to EEBO and ECCO images at reduced rates; it seems probable that they received a discount. However, being offered a discount by the service providers only confirms that the TCP’s scholarly labour had monetary value in the eyes of the commercial entities with which they “collaborated.” [↑](#footnote-ref-56)
56. Beyond the scope of this article is the phenomenon of “crowdsourcing” scholarly labour, especially transcription, which is also often inaccurately framed as a solution to the difficulty and expense of tasks when it simply externalizes that labor. See, e.g., Causer, Tim, Justin Tonra, and Valeria Wallace, “Transcription maximized; expense minimized? Crowdsourcing and editing *The Collected Works of Jeremy Bentham*,” *Literary and Linguistic Computing*, volume 27, number 2, 2012, which suggests that had funds been spent on specialists completing the work directly rather than maintaining a “crowd” infrastructure, the same funding could have produced ten times as many transcripts, except that “no funding body would ever provide a grant for mere transcription alone” (131). [↑](#footnote-ref-57)
57. Since companies and universities are both actually constituted by individual humans, it is of course likely that the people making up both institutions might share values, and even share the goal of furthering human knowledge. However, the value system embedded within an institution are not identical to the values of the individuals who take part in that institution; it is in this sense which I assert that private companies inherently prioritize commercial, rather than scholarly, goals. [↑](#footnote-ref-58)
58. Text Creation Partnership (TCP), “Welcome,” textcreationpartnership.org. Accessed 10 March 2020. [↑](#footnote-ref-59)