# History of the Book Database:

# How Four Repositories of Eighteenth-Century Works Select for Gender

Book history worries about what is lost in a shift to digital books often focus on qualitative, subtle, subjective features of the books themselves: the feel or colour of the paper, and so on.[[1]](#footnote-1) Often, digital facsimiles occlude even very unsubtle information which would be obvious to anyone handling a book in person, such as its length or size.[[2]](#footnote-2) But another form of absence is the book which still sits on its shelf, but undigitized, now worlds apart from its shelf-neighbour which is available digitally. Concerns about uneven digitization, too, are often voiced, though usually in general terms.[[3]](#footnote-3) It seems obvious that there must be systemic absences, but the precise scope and nature of those absences are equally obviously, difficult to define. These absences seem more probable, and more inscrutable, when there are financial/corporate motivations behind the scanning of books, as in the case of Google Books. Google Books is a veil which I have not yet been able to pierce, but it remains so corporate that it is of very little relevance in eighteenth-century studies. More relevant are ESTC, ECCO, HathiTrust, and TCP. Here’s how many works each one has in the 1790s.[[4]](#footnote-4) The TCP thus contains an almost negligible 1% of the titles listed in the ESTC. These four resources cannot have identical holdings, because of their radically different sizes. How different are they? And what is different about them?

Katherine Bode has called for digital histories of literary systems.[[5]](#footnote-5) The system I examine is eighteenth-century studies. This is thus a history of the book database, rather than a history of the book, but with an 18thC lens. Underwood discusses glancingly how, in practice, digital repositories are the foundation of corpus building.[[6]](#footnote-6) Underwood has begun with saying some stuff about sampling.[[7]](#footnote-7) Underwood’s comments about sampling react to Franco Moretti. Franco Moretti famously posed digital methodologies as a solution to the “great unread” of literary history, with distant reading offering a way to finally fully account for the masses of works which ended up on the cutting room floor in the slaughterhouse of literature.[[8]](#footnote-8) This has also been tied to questions about canons, especially since Moretti linked this move to the idea of a ‘world literature.’ But Guillory has already shown us that the canon is not a real list, it is an imagined whole evoked by the partial manifestation of the syllabus.[[9]](#footnote-9) A corpus or a database is not limited the way a syllabus is.[[10]](#footnote-10) The link between canonical representation and digital literary studies is not the database finally offering a capacious enough canon that it can now be ‘opened’ all the way — ie a technological solution allowing literary history to achieve full representation. Instead, doing a history of the book database reveals that the prior for representation is always deferred. Despite the suggestion of completion/wholeness of each individual resource,[[11]](#footnote-11) comparing them to each other immediately reveals that each is only able to *represent*, and attempt to *be representative of*, an imagined larger, prior whole.

Guillory talks about how works are selected from the imagined whole of the canon in order to end up on syllabi[[12]](#footnote-12) — here, works are selected from the real, finite whole of a library’s collection in order to end up in their database. Allen Riddell and Troy J. Bassett note that the books which have been digitized might easily *not* reflect the population of published books, if there has been bias either at the stage of library collection or at the stage of digitization.[[13]](#footnote-13) Give their examples of collection and digitization biases. They found that, indeed, for British novels in 1836 and 1838, women have been underrepresented. This actually *is* something like Guillory’s exclusion hypothesis; these are women who were not barred from access to literacy itself, but who failed to attract the stewardship of future generations. Riddell and Bassett assume that the same general trend will be present in non-British, non-Victorian, and non-fiction literary fields. This research methodology relies on having an actual non-imagined list for the “whole” of “books that were published,” for the proxies to be compared against — Riddell & Bassett use SOMETHING for their list. Expanding the parameters under consideration makes the beginning list harder to find, but is also necessary to test the generalizability of the phenomenon. Remaining within the British context, but expanding to all genres, and to more than one year, I treat the ESTC as, if not exactly a complete list of the whole of “books that were published,”[[14]](#footnote-14) at least a *substantial* list against which other resources can be compared. This is justified for a study of digitality’s impact on the discipline since the ESTC is a reasonable proxy for “books that a scholar could theoretically access in person.” ESTC, ECCO, HathiTrust, and TCP all offer very different access to the imagined prior whole of “English literature of the 1790s.”

## Charlotte Smith in the Digital Archive

Despite the impression they give of always being ‘up to date’, digital resources often reflect much earlier conceptions of that imagined whole. It has now been more than thirty years since feminist criticism started having major successes in 18thC studies. ECCO began in 2003, well after figures like Charlotte Smith had been ‘recovered.’ But the microfilms underlying ECCO date from the 1980s: so ECCO might easily reflect 1980s priorities about what kinds of works are work investing in and redistributing. Charlotte Smith offers an interesting case study in a ‘successful’ recovery project. This allows me to ask: to what extent 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.

Smith’s publishing career began in 1784 and continued until her death in 1806. Our baseline list of “all works published by Smith,” against which we can compare each resource, includes 47 editions of her works published in her lifetime or in the year immediately following her death[[15]](#footnote-15), as indicated in Table 1.[[16]](#footnote-16) None of the four databases includes all 47 of these works. The most interesting exclusions occur earlier and later in Smith’s publishing career. The first edition of her immensely influential *Elegiac Sonnets* (1784), for example, is listed in the ESTC but not available in facsimile anywhere. The publications in the last years of her life, occurring after 1800, 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. Figure 1 shows smaller and smaller databases winnowing down her full output arbitrarily. What, then, is not selected for preservation?

Table

Description automatically generated

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

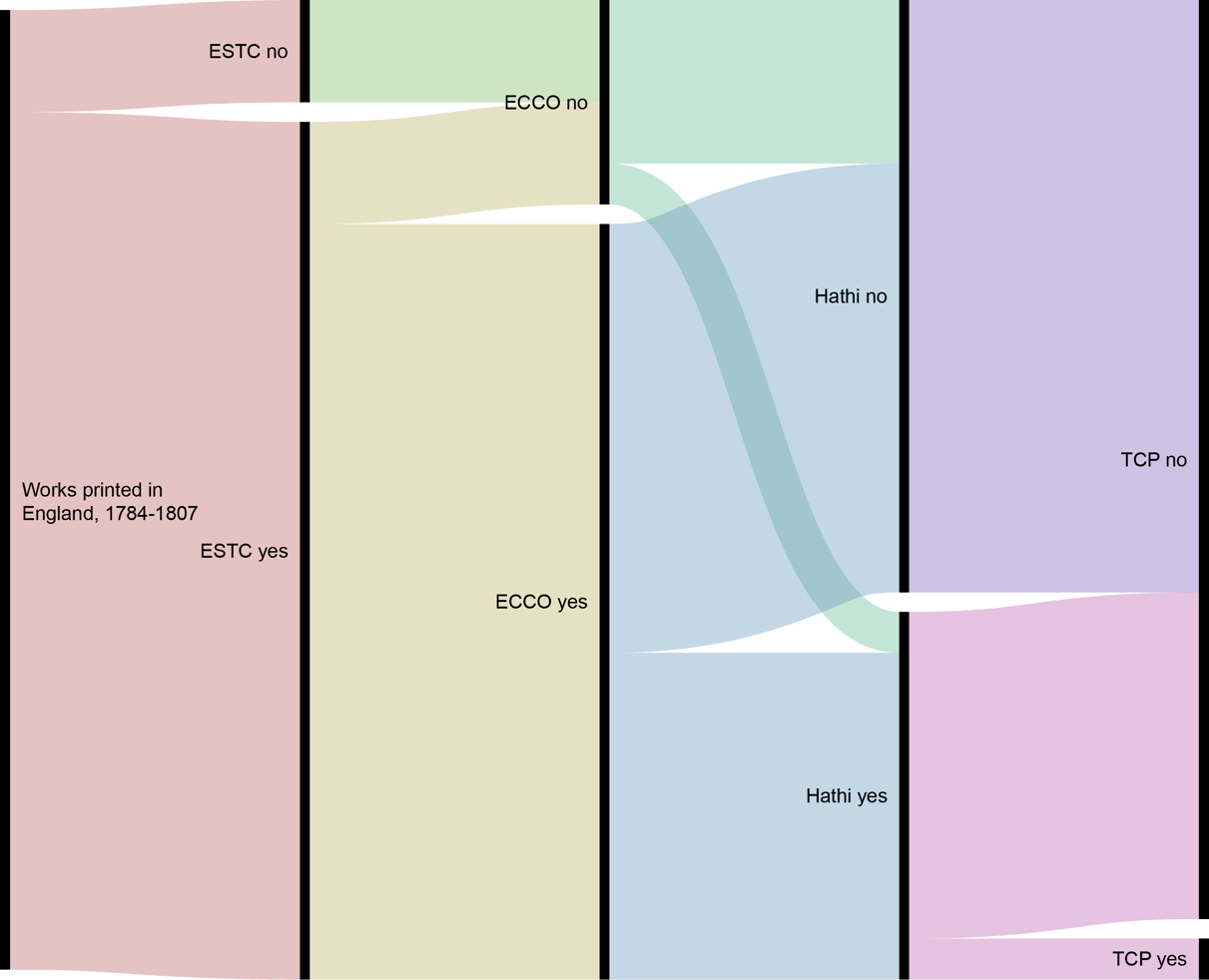


Figure 1: An alluvial chart, showing the winnowing down of Smith’s works from database to database. Of the 47 editions printed in England between 1784 and 1807, 42 are included in the ESTC, and 5 do not appear in the ESTC because they were printed after 1800 and thus fall outside its purview. ECCO contains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

For both ESTC and ECCO, the answer begins with a selection criteria at the site of collection: neither includes any works published in or 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[[17]](#footnote-17) 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.[[18]](#footnote-18) This scanning was undertaken without reference to another bibliography, or any selection criteria over than availability: they want 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 scanned produced from that library’s collection; HathiTrust brings together those libraries’ copies into an academic resource. 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)[[19]](#footnote-19). 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.[[20]](#footnote-20) 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 in other ways, too, the scans making up HathiTrust bear no relation to the scans in ECCO. *The Young Philosopher* (1798), for example,appears in ECCO sourced from a British Library copy, but the HathiTrust images are “Google-digitized” from the New York Public Library. Google’s rapacious book-scanning, unsurprisingly, 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.[[21]](#footnote-21)

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 facilely 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 images of microfilm scans; 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 has been scanned by Google Books and 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,3,68i.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, but the data on page numbers is now misleading.[[22]](#footnote-22) 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.

## 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 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 <$n#fig:databases-venn> 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. I WILL REMOVE PROJECT GUTENBERG FROM THE FIGURE SINCE I DON’T HAVE TIME TO DISCUSS IT HERE. 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; Google Books and Google Ngrams as impossible to discuss except through the academically-inflected proxy of HathiTrust.

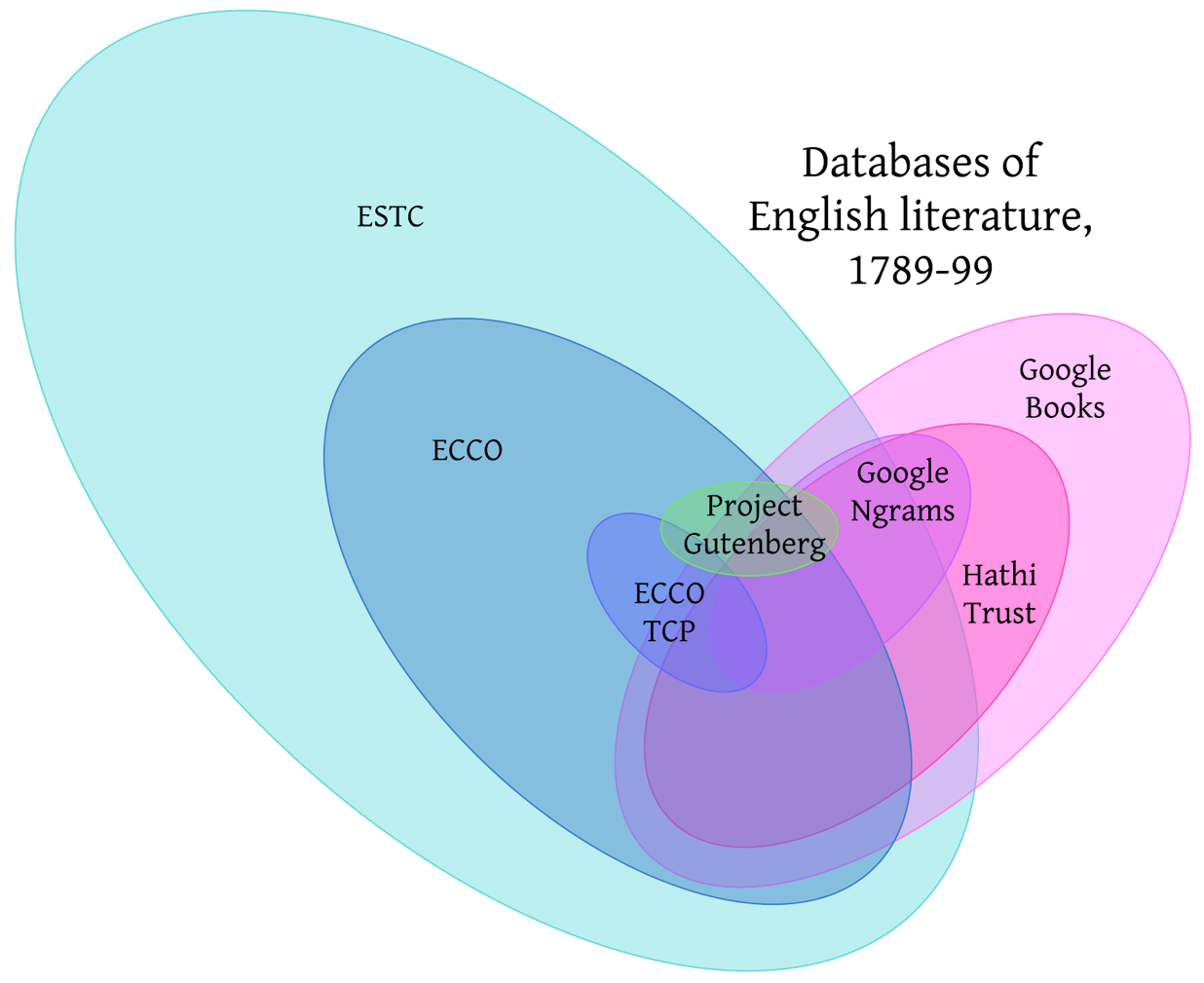


Figure 2: A hand-drawn venn diagram of the relative scale and overlapping holdings of databases containing 1790s literature. The orientation and colour-coding of the ovals groups them into three ecosystems: the academic databases in blue pointing left, the commercial databases in pink pointing right, and the crowdsourced database in green oriented horizontally. I WILL REMOVE PROJECT GUTENBERG FROM THE FIGURE SINCE I DON’T HAVE TIME TO DISCUSS IT HERE.

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, as shown in Table 2 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

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.

1991 End of second edition of Pollard & Redgrave STC, 1475-1640

1991 Exhaustive index to Wing’s STC — after which Bibliographical Society no longer supported Wing

1992 ESTC expanded scope to add serials

1994 ESTC made pre-1700 records available

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 encoding EEBO texts

2000 Project Gutenberg: Charles Franks launches Distributed Proofreaders

2003 ESTC third edition released on CD-ROM

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

2004 Google Print is announced

2005 TCP begins encoding ECCO texts

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

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

2015 EEBO-TCP Phase I books released to the general public

2021 EEBO-TCP Phase II books released to the general public

Table 2: 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).

Some important threads of this history of mass digital archives begins 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, poorly funded and inadequate though they were”[[23]](#footnote-23) throughout the 1950s and 60s pursued the feasibility of systematically accounting for the much larger body of printed work produced in the eighteenth century. Concurrently, Wing’s seventeenth-century STC was underwent 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,”[[24]](#footnote-24) 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. 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.”[[25]](#footnote-25) A pilot project began at the British Library in 1977, under the direction of Robin Alston. Unlike earlier Short-Title Catalogues, which appeared as lengthly print publications, the Eighteenth-Century Short Title Catalogue was conceived as digital from the beginning. As a result, “ESTC records existed in digital form long before many humanists saw computer technology as central to their work.”[[26]](#footnote-26) Robin Alston and Mervyn Jannetta developed their own cataloguing rules, distinct from the Library of Congress MARC and UK MARC standards.[[27]](#footnote-27) 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.”[[28]](#footnote-28)

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.’”[[29]](#footnote-29) Accordingly, the in-progress database was made available through the British Library BLAISE British LibraryAutomated Information SErvice system in 1980 and through the US Research Libraries Group RLIN Research Libraries Information Network system in 1981.[[30]](#footnote-30)

Meanwhile, 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.

1983 also saw the publication of the first subset of the ESTC intended to be accessible beyond the librarians and scholars compiling the file. The ESTC intended to publish editions at particular milestones of completeness, intended for the use of non-librarian scholars. The first, a fiche catalogue and index of the British Library’s holdings, was published in a microform “snapshot” in 1983, but other milestones did not occur according to schedule.[[31]](#footnote-31) The joint Anglo-American interim publication of the ESTC file which was expected to follow on microform in 1984 did not appear.[[32]](#footnote-32) 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,” the ESTC each year seemed to be facing a new budget struggle, and important maintenance labour was several times deferred.[[33]](#footnote-33) 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. To facilitate its use, the ESTC distributed a manual fo searching the online file, “and 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.”[[34]](#footnote-34) 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.”[[35]](#footnote-35) The second volume of the second edition of the STC was published in 1986. In 1987, the Eighteenth-Century Short Title Catalogue expanded its scope to include all materials from the origins of print in Britain to 1800, changing its name to the English Short Title Catalogue.[[36]](#footnote-36)

The 1990s, which saw an explosion of activity fuelled by the popularization of the internet for crowdsourced group-knowledge projects like Project Gutenberg and Wikipedia, were quieter for eighteenth-century academic databases. 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.[[37]](#footnote-37) Meanwhile, the facsimiles making up the *Eighteenth Century Collection* changed hands: by 1997, Research Publications, Inc had become Primary Source Media. In 1998, the Thomson Corporation merged Primary Source Media with two other subsidiaries—Gale Research and the Information Access Company (IAC)—to create a new company called The Gale Group. In 1999, the ESTC officially took over responsibility for hosting and updating STC data.[[38]](#footnote-38)

The 1990s were more more momentous for what would become Early English Books Online. University Microfilms had begun microfilming books in the 1980s as the chief rival for Research Publications, Inc,[[39]](#footnote-39) and by the late 1990s, several thousand reels had been published in two series: ‘Early English Books, 1475–1640’ and ‘Early English Books, 1641–1700’.”[[40]](#footnote-40) University Microfilms became ProQuest, and in 1998, ProQuest began to make available digitised copies of its microfilms across the internet to subscribing institutions — the early days of Early English Books Online (EEBO). The next year, the Text Creation Partnership was formed, to produce human-corrected but machine-readable transcripts of EEBO texts. “The Text Creation Partnership started, in 1999, as a collaboration between the university libraries of Michigan and Oxford, the Council on Library and Information Resources, and the publisher of Early English Books Online, Proquest. The aim was to create high quality ‘standardized, digitally-encoded electronic text editions’ starting with 25,000 titles from Early English Books Online.”[[41]](#footnote-41)

ECCO emerged four years after the beginnings of EEBO, when, in 2003, Thomson Gale (now Gale Cengage Learning) made digital copies of Eighteenth Century Collection microfilms available to subscribers online. But the largest impact on the landscape of digital book repositories was the announcement of Google Print in 2004. Google partnered with major libraries and publishers to gain access to what they estimated at more than 15 million volumes, and set the goal to scan them all.[[42]](#footnote-42) Eighteenth century works were also included, almost incidentally. In 2005, the TCP began encoding ECCO texts, as a somewhat-experimental alternative to the transcripts already produced by Optical Character Recognition (OCR) technologies. In 2006, the ESTC was made available to search for free online, and began adding more information to its listings, to include full title and imprint transcripts.[[43]](#footnote-43) Meanwhile, Google had scanned almost unthinkable numbers of new books. In 2008, however, they begin to face legal repercussions for their “scan first and ask questions later” approach to mass digitization. “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 steward research and cultural heritage by working together than alone or by leaving it to an organization like Google,” says Mike Furlough, the trust’s director.”[[44]](#footnote-44)

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 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 — and also meant, therefore, that they could not display *more* than snippets, even in cases of “copyright orphans.”

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.

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

The naive hypothesis, investigating a historical phenomenon, is that factors like systemic sexism are likely to have an effect. As I have shown in chapter two, the digital archives currently used by eigtheenth-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: little more than the information found on the title page, and a list of libraries where the original can be consulted. 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 systematically opting not to invest 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. 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”[[45]](#footnote-45) — 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.”[[46]](#footnote-46) 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.

I have taken information from four resources most strongly targeted to scholars of eighteenth century literature, namely, the English Short Title Catalogue,[[47]](#footnote-47) Eighteenth Century Collections Online,[[48]](#footnote-48) HathiTrust,[[49]](#footnote-49) and the Text Creation Partnership.[[50]](#footnote-50) Since it would be infeasible in many ways to compare the full holdings of all these resources, we much take a consistent sliver from each resource, and compare within those parameter. Let us examine works published in England,[[51]](#footnote-51) between 1789 and 1799.[[52]](#footnote-52) After removing misclassified works, 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. At this scale, let us follow the fate of a larger figure than Charlotte Smith herself, and see how “the named woman” fares.[[53]](#footnote-53)

Smaller, less comprehensive archives do generally dedicate a higher percentage of their space to male authors. As Figure 3 shows, 43% of the 51,090 titles in the ESTC from this decade are identifiably by men,[[54]](#footnote-54) which rises substantially to 69% for HathiTrust and 67% for the TCP.

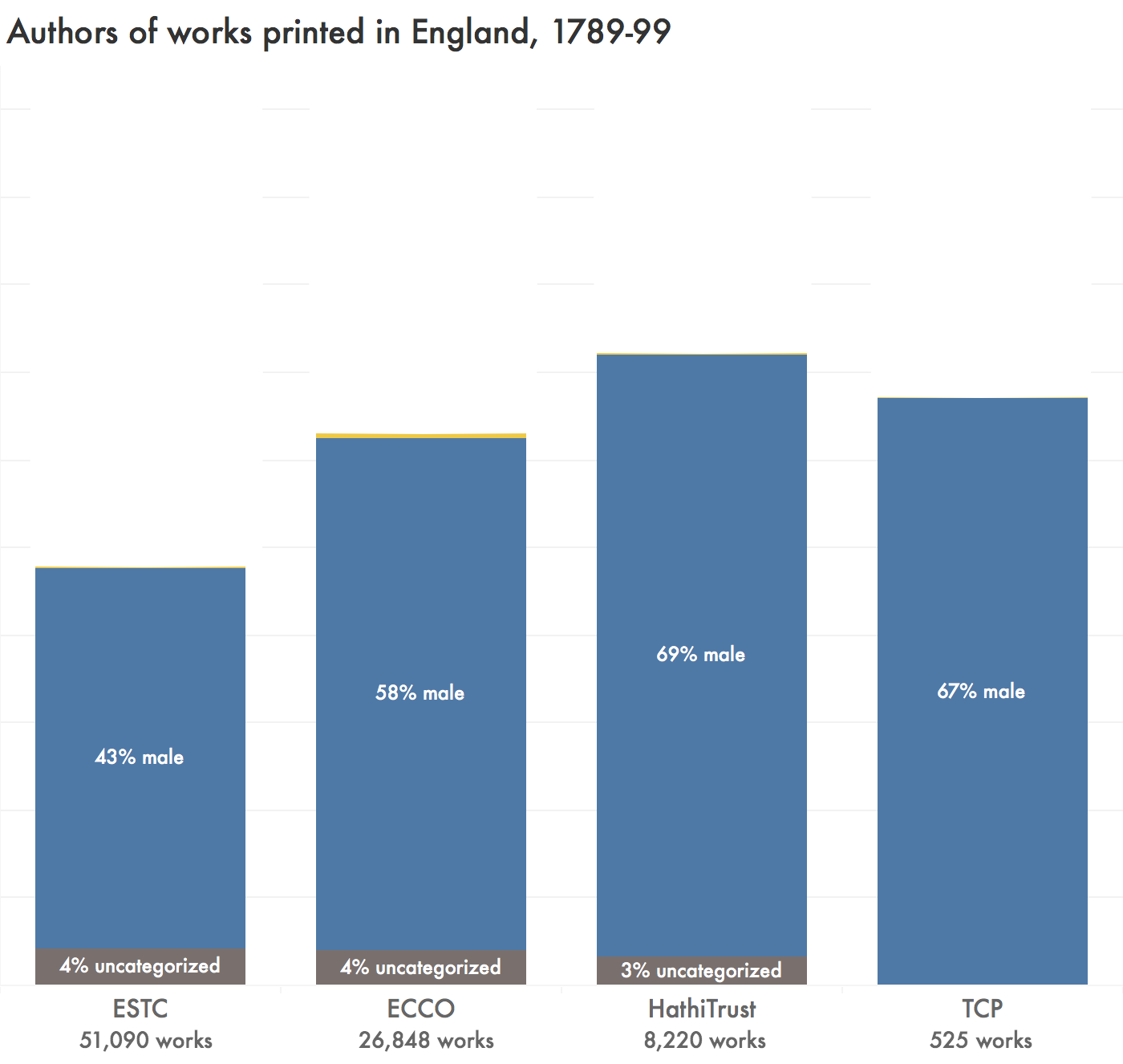


Figure 3: The percentage of works published in England 1789-99 with male authors, as identified by the ESTC, ECCO, HathiTrust, and the TCP.

However, the increase in male authors is not at the expense of identifiable male authors, who also make up a stable or increasing percentage of the works — 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.[[55]](#footnote-55) The ESTC and ECCO have a lot of medical texts, sermons, natural histories, legal opinions, and other genres from which women are de factor excluded. HathiTrust’s slight increase in female authorship might show a collections-level bias where university libraries prefer slightly more interesting or 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. The TCP is the one operating based on the most recent decision-making (since even HathiTrust is constrained to the library collecting decisions of past decisions, and thus doesn’t necessarily reflect the priorities of 2000s

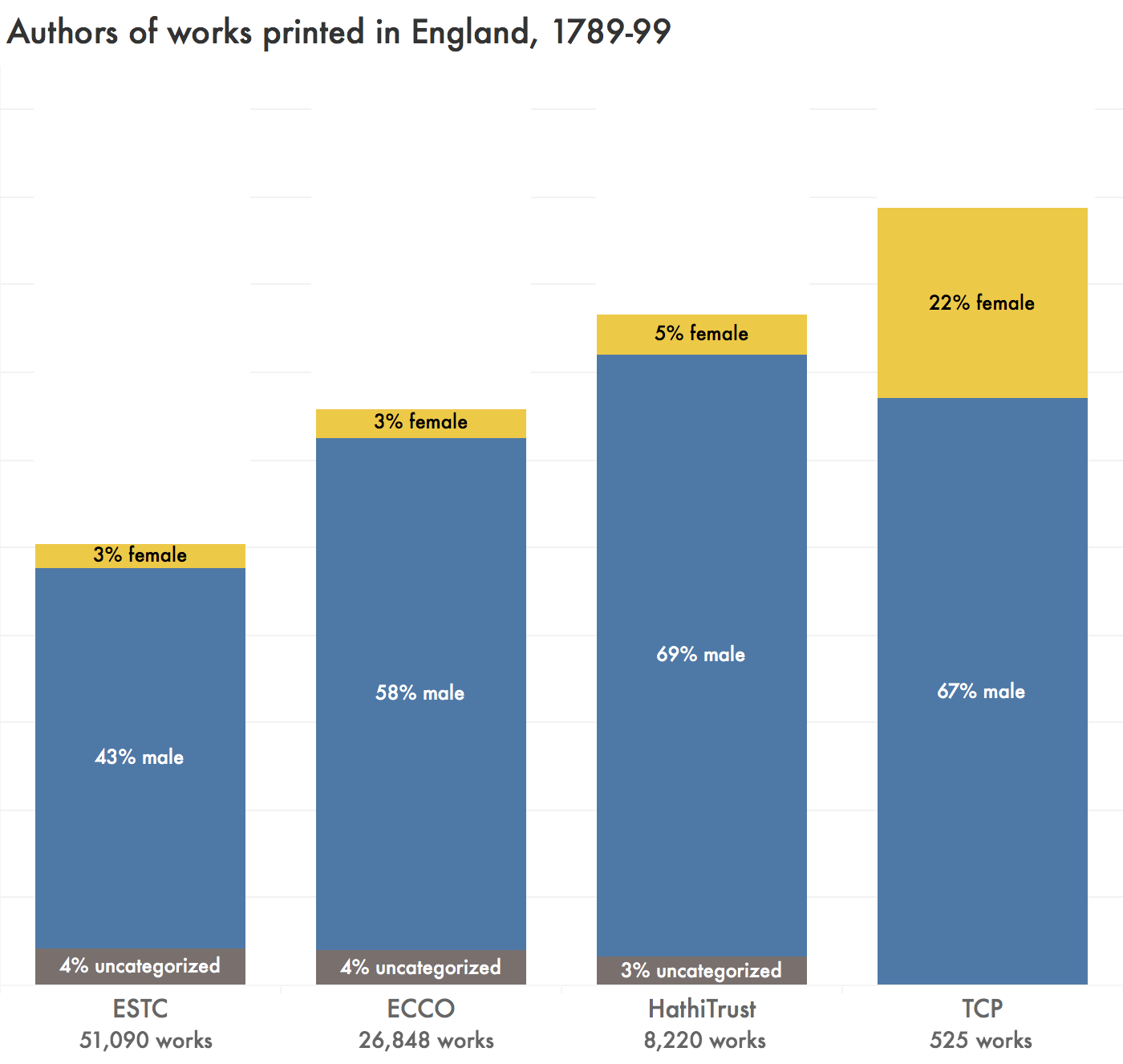


Figure 4: The percentage of works published in England 1789-99 with male versus female authors, as identified by the ESTC, ECCO, HathiTrust, and the TCP.

In Figure 5, however, we can see that 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.”[[56]](#footnote-56) 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. “Unsigned” or “pseudonymous” people are not an identity category in the real world requiring equitable social standing: the unsigned and pseudonymous authors, if identified, would all turn out to have names, genders, races, class standings, and all the other traits of humans in the world. What their status as “unsigned” in the database really means is not that those traits are absent — it means that those traits are undetermined at the site of the title page.

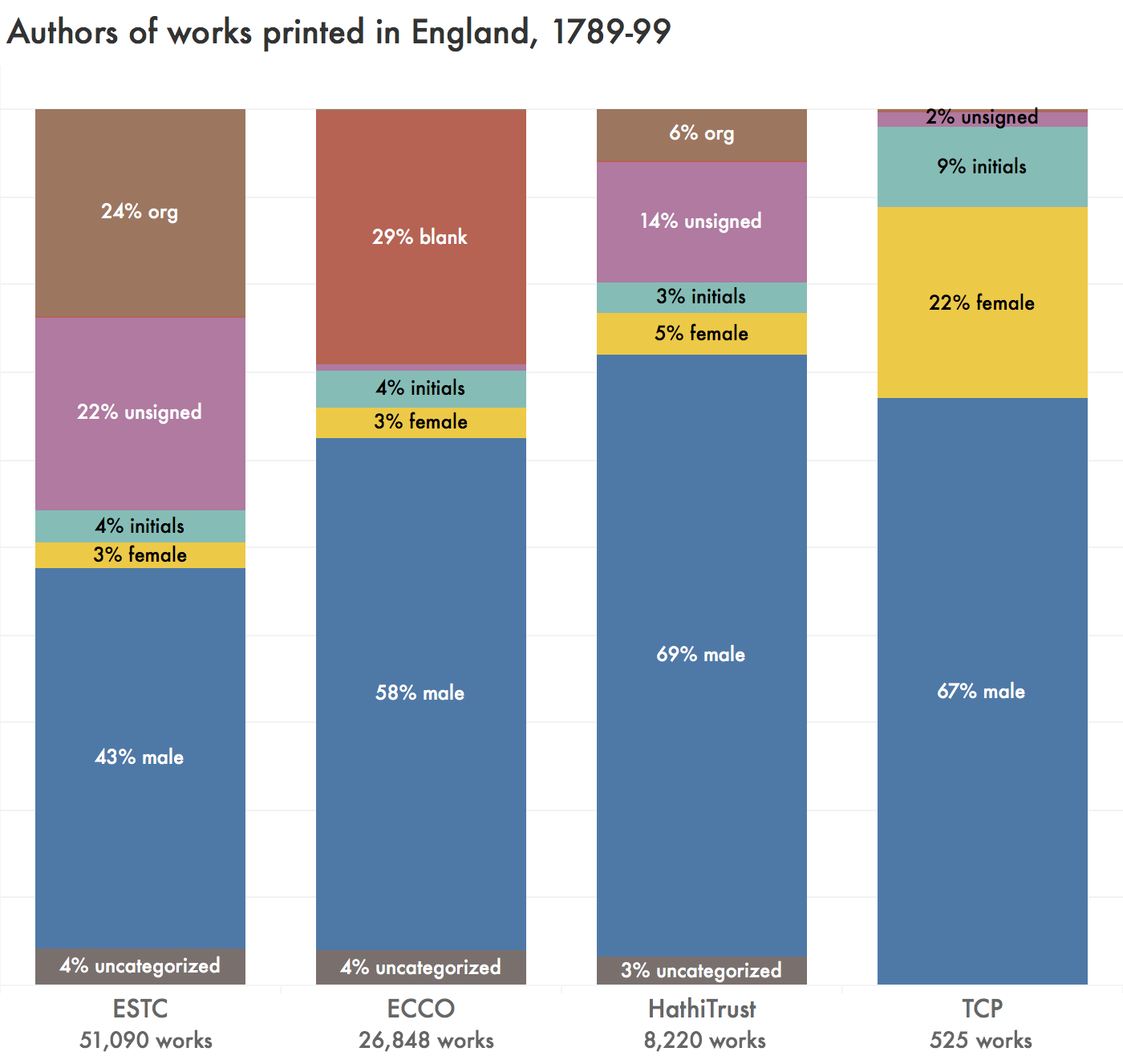


Figure 5: The relative authorship status of works published in England 1789-99, as identified by the ESTC, ECCO, HathiTrust, and the TCP. Works by “org” are attributed to organizations as their authors, such as the House of Commons. “Unsigned” works have no information in the author field. “Blank” in ECCO captures both unsigned works and works by organizations, which I have not yet been able to distinguish. Works by “initials” include those by, e.g., W.H. Ireland, even when the database has provided a clearly gendered expansion of the initials (e.g., a listing of “W[illiam] H[enry] Ireland”), to reflect the fact that the original author listing was of ambiguous gender. From my casual inspection, in nearly all cases of attributed to initials where a known author name is also supplied, the author is male. 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; anecdotally, these appear to consist mostly of pseudonyms, mostly male.

There are two ways to explain the disappearance of unsigned authors from smaller corpora. The first explanation is 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 that the further 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” here, they need not only be absent from the title page, but they also need to have avoided a later scholarly attribution of authorship. Names are placed into the relevant database fields wherever possible, supplying new information that cannot be found on the title page and may have been entirely unavailable to eighteenth century readers. 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 in the internal database, and in the data I was able to access. 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 the particular implementation also occludes how eighteenth century readers actually encountered author information. An identified name overwrites the information recorded on the title page.

Scholars have worked to infrastructurally eliminate the “unsigned” author. “Real” names are provided even in cases where this substitution contains little real information. One work, titled “Observations on the origin and effects of the Test Act” and printed in 1790, has its author listed in the ESTC as “Hudson, active 1790.” A note suggests that this attribution is somewhat shaky: “Attribution from Halkett & Laing which seems to be based on a MS attribution on O copy title page, no longer legible. O copy with MS date May 18. 1790. Verify number and position of advertisement leaves.” This is the only item in the ESTC attributed to “Hudson, 1790” so identifying “Hudson” in this way does not provide value by linking the work to another. In contrast, the title page attribution identifies the author as “a dissenter” — a more informative identification, and even one which might form the basis of a valuable database query. How many works are attributed only to “a dissenter,” and what do those works in common? The existing database infrastructure makes it highly challenging to answer those questions.

## “Representativeness” in Databases

Katherine Bode, in “The Equivalence of ‘Close’ and ‘Distant’ Reading,” argues that Franco Moretti and Matthew Jockers replicate the approaches of New Criticism with their corpora, and calls for “a new scholarly object of analysis”[[57]](#footnote-57) that directly examines historical and textual context of corpora as representations of “literary systems.”[[58]](#footnote-58) 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.”[[59]](#footnote-59) 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.”[[60]](#footnote-60) 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.”[[61]](#footnote-61) 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, on the surface of it, is not without grounds. 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. 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 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.[[62]](#footnote-62)

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.”[[63]](#footnote-63) 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;”[[64]](#footnote-64) 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.”[[65]](#footnote-65) 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. 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 can also render these biases no longer unavoidable.

## Conclusion

Women made up a nearly negligible portion of the titles published in England during the 1790s. If one were to choose a title at random, it would be more likely to be written by a man named John than by a woman. But we do not choose titles at random— even in large-scale digital humanities research. We neglect the genres from which women are most strongly excluded: medical texts, legal texts, sermons, and other discourses tied to specialized (male) occupations. We prefer fiction and poetry, the forms most open to women. Even then, we overrepresent women. I want to be very explicit that the overrepresentation of women is not inappropriate: it strikes me as a very positive sign that we might have grown less sexist in the intervening two hundred years. Instead, what this work shows is that the logic of “representation,” which underpins discussion of literary canons and syllabi, is not a good fit for understanding the infrastructural impact of contemporary digital databases.

1. There must be several people to ref here — start with Ian Gadd [↑](#footnote-ref-1)
2. I read something about duodecimos and folios, find and ref it [↑](#footnote-ref-2)
3. See, for example….. several people to ref/quote here [↑](#footnote-ref-3)
4. These figures, and all others discussed here, are from samples I acquired from the four resources, and are likely to differ slightly from database queries that are run today. My ESTC data is from DATE; ECCO data from DATE; HathiTrust data from DATE, and TCP data from DATE. [↑](#footnote-ref-4)
5. cite Bode [↑](#footnote-ref-5)
6. cite Underwood [↑](#footnote-ref-6)
7. cite Underwood [↑](#footnote-ref-7)
8. cite Moretti. also talk about how the field is moving away from Moretti; cite Klein and Bode [↑](#footnote-ref-8)
9. cite Guillory [↑](#footnote-ref-9)
10. ref Underwood, who comments on this; also talk about how I’m extending/responding to his comment [↑](#footnote-ref-10)
11. quote that guy who talks about how ECCO makes it seem like anything that’s not in ECCO just doesn’t exist [↑](#footnote-ref-11)
12. cite guillory [↑](#footnote-ref-12)
13. cite Riddell and Bassett [↑](#footnote-ref-13)
14. For example, the ESTC intentionally excludes playbill, and surely unintentionally excludes many other titles. [↑](#footnote-ref-14)
15. I include the first year after her death to capture publications which Smith herself prepared for publication, such as *Beachy Head*, but no editions pat 1807 in order to focus this experiment on Smith in her own time, rather than Smith’s ‘afterlives’ or reception. [↑](#footnote-ref-15)
16. This table synthesizes bibliographic information from several works on Smith, including LIST THEM. [↑](#footnote-ref-16)
17. Later known as Primary Source Microfilm, an imprint of the Gale Group, . [↑](#footnote-ref-17)
18. Although Google Books’ scanning hit its peak in YEAR, and they are very quiet about how much scanning they are doing (they don’t update their own Google Books history page), new scanning is indeed ongoing, as seen in XXX press release. [↑](#footnote-ref-18)
19. 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-19)
20. 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 libraries (with no indication of which library provided it). [↑](#footnote-ref-20)
21. Both works were first reproduced in the microfilm version produced 1982-2002 in by Research Publications, 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. [↑](#footnote-ref-21)
22. 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-22)
23. Korshin 209 [↑](#footnote-ref-23)
24. Vander Meulen 268 [↑](#footnote-ref-24)
25. Crump 105 [↑](#footnote-ref-25)
26. Karian 283 [↑](#footnote-ref-26)
27. Korshin 211 [↑](#footnote-ref-27)
28. Crump 105 [↑](#footnote-ref-28)
29. Vander Meulen 270 [↑](#footnote-ref-29)
30. Norman n. pag. [↑](#footnote-ref-30)
31. Crump 105 [↑](#footnote-ref-31)
32. Korshin 212 [↑](#footnote-ref-32)
33. Alston n. pag. [↑](#footnote-ref-33)
34. Crump 106 [↑](#footnote-ref-34)
35. Tabor 367 [↑](#footnote-ref-35)
36. Norman n. pag. [↑](#footnote-ref-36)
37. Vander Meulen 269 [↑](#footnote-ref-37)
38. Vander Meulen 270 [↑](#footnote-ref-38)
39. The competition between University Microfilms and Research Publications, Inc in the 1980s is, in fact, the explanation for many current technological oddities around the difficulty of using EEBO and ECCO in concert. Although both EEBO and ECCO are closely tied to the ESTC in similar ways, and serve similar functions on either side of the arbitrary chronological dividing line of the year 1700— all of which would seem to make cooperation between the two resources a natural choice— they remain owned by rival companies which place them in competition. [↑](#footnote-ref-39)
40. Gadd [↑](#footnote-ref-40)
41. Gregg n. pag. [↑](#footnote-ref-41)
42. It is hard to resist the drama of anecdotes about the early days of what would become Google Books. “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) [↑](#footnote-ref-42)
43. “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” (Tabor 367) [↑](#footnote-ref-43)
44. Van Helden n. pag. [↑](#footnote-ref-44)
45. Guillory p 8 [↑](#footnote-ref-45)
46. Guillory 15 [↑](#footnote-ref-46)
47. 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. They delivered a dataset containing 51,965 records, matching a database query for works published in England 1789-99. I conducted additional data cleaning in OpenRefine to check all publication dates and locations, removing any which fell outside my strict parameter (e.g., those attributed to a range of dates when the earliest date preceded 1789), which brought the total number of records analyzed down to 51,090. [↑](#footnote-ref-47)
48. MARC records of ECCO holdings were generously provided by the University of Toronto Libraries. These records were not straightforward to work with; I used MarcEdit to convert them to csv format, then OpenRefine to remove the many duplicates (based on the unique Gale Document Number) and then to remove all works not published in England 1789-99. [↑](#footnote-ref-48)
49. HathiTrust data is publicly available. I created a HathiTrust “collection” of works published in England 1789-99 and downloaded the related metadata. [↑](#footnote-ref-49)
50. 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-50)
51. 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. “Plymouth, M.A.,” therefore, would be removed as referring to the city in the United States, but “Plymouth” alone would be retained as potentially referring to the city in Devon. Some of these assumptions were a bit improbable, as in the cases where I assumed “Halifax” referred to the township in West Yorkshire rather than to the city in Canada, but I preferred to employ a consistent systematic rule (“England until proven guilty”) rather than follow a case-by-case hunch about likelihood. I do no expect the 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 strongly different outside of England. [↑](#footnote-ref-51)
52. 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-52)
53. After discovering the limits of multiple automatic methods for guessing gender from a first name, all of which were clearly not created with the eighteenth century in mind, I assigned the gender assessments manually. Most texts were associated with a small number of extremely common names which presented no interpretive difficulties (John, William, Thomas, James), but as I reached the long tail of publications I looked up all necessary specific individuals to confirm that, for example, Tench Coxe was male and Benedikte Naubert was female. The methodology of assigning first names themselves as “male” or “female” names is not one I would endorse for all historical periods (it is clearly untenable in our present moment), but in my very close encounter with the specific data of 1790s names, the assumption of a strong gender binary as well-matched by eighteenth century social practices. [↑](#footnote-ref-53)
54. 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-54)
55. I have calculated this figure from the table of data 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, 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. Page 47. [↑](#footnote-ref-55)
56. The category “unsigned” here captures works to which no name at all was signed (essentially, anonymous works, though in this period of course the absence of a name does not imply the concealment of a name as it would in the contemporary definition of “anonymous”) as well as pseudonymous works which offered no clues to an author’s gender (e.g. “A friend of peace”). “Organizations” include missives from various branches of the government, reports from scholarly and charitable organizations, catalogues from various companies, and other works which are easily rejected as “non-literary.” The identifier “blank,” for ECCO, appears to capture both unsigned works and works by organizations. ECCO so thoroughly rejects the idea that “Society for the Improvement of Naval Architecture” or “Benevolent Institution for the Sole Purpose of Delivering Poor Married Women at Their Own Habitations” could be considered as “authors” that it does not record them in the “author” field. [↑](#footnote-ref-56)
57. Bode “Equivalence” 79 [↑](#footnote-ref-57)
58. Bode “Equivalence” 97 [↑](#footnote-ref-58)
59. Laura Klein (n. pag.) [↑](#footnote-ref-59)
60. Bode “Equivalence” 78 [↑](#footnote-ref-60)
61. Klein n. pag. [↑](#footnote-ref-61)
62. Bullard 756 [↑](#footnote-ref-62)
63. *World of Fiction,* 47. As an example of the frustrating impacts of the need to brand services as universally useful: I attended a daylong university workshop about the new Gale Digital Scholars Lab in the company of a medievalist who arrived, discovered that no medieval materials were available in the Digital Scholars Lab, nor even the Early English Books Online collection, and spent the morning awkwardly listening to speakers effuse about the incredible scope of U of T’s subscriptions. [↑](#footnote-ref-63)
64. TCP, “About” [↑](#footnote-ref-64)
65. TCP “Welcome” [↑](#footnote-ref-65)