The Anatomy of Eighteenth Century Collections Online (ECCO)

Mikko Tolonen, Eetu Mäkelä, Leo Lahti

Eighteenth-Century Studies, Volume 56, Number 1, Fall 2022, pp. 95-123 (Article)

Published by Johns Hopkins University Press
DOI: *https://doi.org/10.1353/ecs.2022.0060*

The Anatomy of Eighteenth Century Collections Online (ECCO)[1]

Mikko Tolonen, Eetu Mäkelä, and Leo Lahti

In less than two decades, *Eighteenth Century Collections Online* (ECCO) has become the primary digital source for the study of printed eighteenth-century texts in the English language.[2] Its relevance to research on the history of literature, however, goes well beyond eighteenth-century studies. Soon after its launch ECCO was called revolutionary, something that every self-respecting university must have. Hopes were high that it would change the ways in which eighteenth-century scholars are able to reveal cultural trends, followed, of course, by more skeptical voices noting, among other things, its hefty price tag.[3] While different debates around the dataset have been many, surprisingly little has been said about the actual content of ECCO. We want to change this. The main aim of this paper is to offer an account of the anatomy of ECCO: what is missing, what are the imbalances, and how representative it is with respect to its source catalog. We use the metaphor of

Mikko Tolonen is an Associate Professor in Digital Humanities at the University of Helsinki and the leader of the Helsinki Computational History Group (COMHIS). His part of this work has been funded by the Academy of Finland under Grants 1333716 and 1347706. He'd like to thank also Helsinki University Library for funding the Open Access of this article and members of COMHIS for supporting the work.

Eetu Mäkelä is an Associate Professor in Human Sciences-Computing Interaction at the University of Helsinki. His research group seeks to figure out the technological, processual, and theoretical underpinnings of successful computational research in the humanities and social sciences. He thanks the Academy of Finland for funding this research (Grant 347709).

Leo Lahti is an Associate Professor in Data Science (Computational Humanities) at the University of Turku. His research team focuses on the theory and methods in the analysis of complex natural and social systems. He thanks the Academy of Finland for funding this research (Grant 348946).

anatomy to underline the evolving nature of ECCO and different layers of interconnectedness in its content. Our interest is its make-up and composition, not its layout, organization, or other technical aspects of the database.

The success of ECCO makes sense when one considers that it is not a standalone product invented in the late 1990s and realized in 2003, but arose from a longer process of curation. The most important basis for ECCO is the eighteenth-century cataloging project from the late 1970s, renamed the English Short Title Catalogue (ESTC) in 1994 after expanding to other periods.[4] This metadata catalog currently includes bibliographic information for over 480,000 documents that were published between 1473 and 1800. Most of the documents included in the ESTC (and also in ECCO) were written in English and published in the British Isles and North America, collected from hundreds of libraries worldwide in a union catalog coordinated by the British Library.[5] After the ESTC had been launched, "the Eighteenth Century" microfilming project followed in 1981. ECCO was born when microfilms of that project—originating mainly from the British Library, Oxford, and Cambridge—were scanned in 2000–2002, and optical character recognition (OCR) was applied to associate the pages with automated transcriptions of the texts. Gale then published these texts in a web interface in 2003 that also enabled text search of the transcriptions.[6]

Remarkably little has been written about what material ECCO does and does not contain. This point is raised by Cassidy Holahan in her recent essay "Rummaging in the Dark: ECCO as Opaque Digital Archive."[7] Upon ECCO's release, Gale claimed that the dataset "contain[ed] every significant English-language and foreign-language title printed in the United Kingdom between the years 1701 and 1800." In truth, however, it has always been a resource in motion.[8] Following its initial release of around 135,000 documents in 2003 (termed ECCO1 in this article), some 47,000 further titles were released in an expanded version in 2009 (termed ECCO2 in this article). Gale attributes the ECCO2 addition to the fact that the ESTC expanded after ECCO1 was released with "materials previously unavailable, undiscovered or inaccessible." Thus "close to 7 million pages" were added, including "works previously too fragile to be handled, owing to rapid developments in scanning technology."[9] In the near future, based on conversations with Gale representatives, there will be further materials added by Gale to produce ECCO3 (a term used in this article).

Given the changing nature of digital data collections such as ECCO, rigid demands that collections must include "everything" seem somewhat naive.[10] As David McKitterick points out, it is in the nature of such collections that new material accumulates over time.[11] In addition, what is in ECCO is shaped by its long history. It is missing a sizable chunk of publications from the United States that are recorded in the ESTC, for example, as we shall show. This shortcoming may originate from the fact that the Eighteenth Century microfilm series that later formed the basis of ECCO competed with other microfilming projects for the period, including one focusing on imprints from the United States. In other words, certain material is missing from ECCO when compared to the ESTC because at some point it was available elsewhere on microfilm.

The problem with assessing ECCO's representativeness has been that usually users only have limited materials available. Our access to full-text in ECCO

and the ESTC metadata enabled us to design and implement our own research ecosystem without being at the mercy of the interfaces.[12] In this first large-scale analysis of the imbalances in ECCO we use the ESTC as the point of reference. This interdependence between the content of ECCO and the ESTC is also the most important reason to talk about its anatomy. Although not in itself a wholly accurate proxy for all publishing activity at the time, the ESTC is still the best and most comprehensive source available.[13] Furthermore, even if the ESTC is not completely representative of all the English language material published over the course of the eighteenth century, by comparing it with ECCO we are still able to produce a better picture of the dynamics of both collections. This, in turn, could inform the reasoned use of each resource.

In its dominant mode of use, the ECCO interface—much like the "digital humanities" more broadly—functions as a tool enabling faster and more efficient retrieval of information about particular topics.[14] One particularly important context for the use of ECCO is teaching.[15] All in all, most users are looking for accurate "hits" while searching for particular strings of characters of authors who interest them, keywords or quotes from their favorite works, for example.[16] If we treat ECCO naively, we are at risk of emphasizing certain authors or works in comparison to others. It is important that researchers avoid the vicious circle of unintentionally focusing attention on already overrepresented aspects of ECCO. At the same time, the reliability of keyword searches is a major concern among the main group of users.[17] Indeed, in the early 2010s, the inadequate rendering of all the texts through OCR in ECCO caused a plethora of worries about the potentially severe shortcomings of keyword searches and other ways of examining the digitized sources in the collection.[18]

As an example, in her survey of how ECCO has been used for eighteenth-century scholarship, Holahan notes that ECCO has been used for various argumentative purposes to compare how often the phrases "creative genius," "breeches," "creativity," "system," and "young invader/young pretender" appear in the corpus for different years. For the sake of argument, let us say that the frequency of one of these terms drops in the later part of the eighteenth century. On the face of it, we might consider this to mean that the topic has gone out of fashion. Now, consider—as we will show—that a much larger share of late-century societal pamphlets is missing from ECCO as compared to earlier pamphlets. Could the frequency drop have been caused by this imbalance in the data instead of the topic going out of fashion? Or what if the frequency in fact increases when these omissions are accounted for? In truth, such an increase may be due merely to the explosion of publishing activity, which also feeds into ECCO, causing the latter half of the century to contain many more publications than the first half.

For those using ECCO keyword search interfaces, the main problem caused by representational and OCR quality imbalances is the inability to reason about whether something is missing due to not having existed in the first place or merely because the dataset is skewed. Anyone perusing a traditional physical archive would at least get an overview of it and its coverage as a whole, whereas what one sees in ECCO is only the matching documents along with a vague promise that the digital archive contains "nearly everything" that is important.

This issue of representational balance is even more relevant to those trying to use ECCO quantitatively.[19] In short, if one draws quantitative inferences from a dataset, it is important that the dataset's composition accurately represents the aspects of reality for which it is used as a proxy. If it does not, any inferences drawn from it are liable to be severely misguided.[20] This is a crucial concern at a time when more and more people are using ECCO as a dataset to make claims about linguistic changes over the whole eighteenth century.[21] For example, many historians are eager to explore conceptual and semantic change based on large historical data.[22] As another example, the Linguistic DNA project aimed to use ECCO as one of the main sources to uncover "the DNA" of historical English, although the research team ended up working with Early English Books Online - Text Creation Partnership (EEBO-TCP) that is manually keyed, mostly error-free data.[23]

At Helsinki Computational History Group (COMHIS), we have been working on different aspects of the ESTC and ECCO for the past decade. The overall strategy is to enable bibliographic research and full-text mining on available early modern British data sources.[24] When taken as a serious scientific effort, this is the kind of work that develops iteratively. The relevance of interoperability is underlined by Bullard.[25] Currently, we have advanced to a stage where we are able to work on projects using High Performance Computing to combine bibliographic metadata and the development of large language models to define boundaries of historical discourses over time.[26] The question of representativeness of the data sources that our group uses is perennial to all aspects of our own work.

In other words, what people working for Gale referred to a while ago as "unexpected ways" of using their data is now turning out to be the norm when researchers hope to move beyond keyword searches towards systematic large-scale analysis.[27] Recently, Gale even invested in something called the Digital Scholar Lab to enable the use of digital humanities tools for their data and to make such research easier.[28] For all these reasons, the representativeness of ECCO should be of high concern to all of us studying the eighteenth century.

In the rest of this essay we will shine some light into the darkness identified by Holahan by offering a *de facto* comparison of the contents of ECCO and the corresponding records of the ESTC. Such a comparison can then function as a starting point for further reflection on the intentional features of the two collections as databases.[29] It is not our purpose to bash ECCO for missing certain material, or even to delve deeply into why certain categories of works are over- or under-represented. We merely wish to help users understand ECCO as it currently exists and to aid them to think about what consequences its representativeness might have for their own research. To us the most important aspect of the composition of ECCO is a question of balance rather than total coverage: whether those using ECCO will find the same balance of viewpoints, opinions, and contents as they would if they consulted archival collections.

DATA HARMONIZATION AND QUANTITATIVE COMPARISON

Bibliographies such as the ESTC contain rich information on book printing.[30] According to Robin Alston, however, statistical estimates undertaken in the

late 1980s indicate that there were substantial errors in the data.[31] There have been major improvements in ESTC coverage since Alston's time due to increased contributions from institutions in the United Kingdom, the United States, Europe, and around the world, but it should be kept in mind that the ESTC itself is a changing artifact. Library catalogues have been designed for information retrieval from individual documents rather than for comparative large-scale analyses. Hence, in general, the entries in library catalogues have to be substantially harmonized and curated before this data can be reliably used to assess variations in book-printing activities.

Examples of common issues that we have dealt with in COMHIS with the ESTC catalog include disambiguation of alternative name variants (e.g., for authors, publishers, and places), different entities that share the same name and have to be distinguished, resolving errors and inconsistencies encountered during the semi-automated data curation, and recovering missing information from complementary sources. In addition, we have done dedicated interpretation for some information fields: the fields of physical extent and physical dimension, for instance, follow the Machine-Readable Cataloging (MARC) notation and had to be converted to numeric page-count information by means of rule-based analysis. We have further enriched the data in the library catalogs by incorporating additional information about authors, publishers, and place of publication. The full technical details of these harmonization and curation efforts for the ESTC catalogue are extensive, and instead of repeating them here we refer to our recent work that provides a more comprehensive description.[32]

Hugh Amory claims that it is impossible to use the ESTC as a data resource for statistical analysis.[33] This has been sensibly countered by others, and we have demonstrated that bibliographical records can indeed be used for statistical purposes.[34] Moreover, book historians based their earlier rudimentary statistical approaches to the development of British printing on the ESTC.[35] In terms of representativeness, what should be understood about any catalog or tool built upon it is that all collections are limited to a specific scope that has to be accounted for in the analysis. Despite the limitations, historical data is useful for modeling and enhancing understanding of the past when the data-generation process is properly understood and addressed.

Regardless of the benefits of automated data processing and curation, however, one needs to remain critical about the overall representativeness of the data. Smaller books may be more easily lost over time, general collections are often complemented with specialty collections, and certain types of information such as that about the publisher and the author may be more likely to be missing, or more challenging to interpret, for certain time periods.[36] Digitization and converting catalog entries and full texts into machine-readable formats create a set of problems in themselves, such as the dependency of users on search functions that "hide the catalogue." To illustrate this point, Amory provocatively claims that the ESTC resembles "the Holy Roman Empire" because "it is neither English, Short-Title, nor a Catalogue, since the 'cataloguing' is only a response shaped by the system at the user's request."[37] However, joint analyses of multiple, independently collected catalogs could highlight broad patterns that are systematically observed across different collections, and thus support the critical analysis of historical trends.

There will always be under- and over-represented parts in any data put together from earlier historical collections. It is clear that there are gaps in the ESTC as well, for example. But what is significant for the purpose of this article is that a comparison between ECCO and the ESTC makes sense because the ESTC is the backbone of ECCO: they are not independent data collections. What is of major importance is being able to estimate these imbalances and levels of representation based on subject knowledge and statistical expertise. Understanding the anatomy of ECCO through the ESTC is particularly appealing because, in principle, they each make a good base for the quantitative study of book production due to the relative technological uniformity of the hand-press era they cover. This, together with our efforts to harmonize the ESTC, mitigate many of the limitations mentioned above.

THE MAIN FINDINGS

The differences in composition to which we pay particular attention in this article relate to publication format, reprinting phenomena, the overrepresentation of popular authors, and the underrepresentation of certain publication places in ECCO, as well as how many of these are conditioned on the year of publication of the work.

The Relevance of the document type: the focus in this section is on the differences between publications of different types. The analysis takes into consideration publication length as well as differences between gatherings (e.g., folio, octavo). We show that the representation of pamphlet-sized documents in ECCO is poor overall, but particularly after 1770. Although there is a 10 percent linear decrease in representation for book-sized objects over the course of the eighteenth century, there is a sharp drop for pamphlets, particularly governmental documents. The more frequent inclusion of pamphlets in ECCO from authors that are also otherwise well-known is akin to the rich-get-richer phenomenon, also known as the *Matthew Effect*: if you have published just one book, for instance, the chance of having your pamphlet-sized documents included in ECCO increases by 30 percent.[38]

Places of publication: this section broadens aspects of ECCO related to the temporal and publication format by including spatial aspects in the discussion. We show that although, as expected, London dominates the data, there is a clear country-specific effect that should be taken into consideration. This is best recognized by comparing products printed in Ireland and the United States. The overrepresentation of Ireland and the underrepresentation of the United States applied to both book and pamphlet-sized documents. Regional variance also changes over time, as in the case of Scotland. We also analyze the representativeness of different cities that gives us more detailed information of the overall country trends, especially in the United States.

Editions and reprints: the relevance of reprints in ECCO has so far been understudied and largely neglected in terms of data analysis. What we show is that first printings are better represented than reprints or singly printed works in all the different analytical categories. In short, most books are reprints whereas most pamphlets are single printings. First editions of works that are reprinted are better represented by a median of 10 percent than either their reprints or works that are published just once: 60 percent of the books and 25 percent of the pamphlets

included are reprints; the median age of book reprints is twenty years, and the median age of pamphlet reprints is ten years.

Authors: this section expands the analysis by focusing on known authors in ECCO and paying particular attention to authors/writers of literature and religion. One of the most common ways of using ECCO is to take an author-centric approach and to look at knowledge production through the lenses of particular authors, which is why we extensively examined the most popular authors in ECCO. We also selected popular but poorly covered authors for analysis, looking at works, editions, books, and pamphlets from them. This examination reveals considerable differences in the representation of individual authors in ECCO. In our author-centric way of grasping cultural change, it makes a substantial difference if we inadvertently place relevance mainly on overrepresented authors. We also analyze female authors based on the same categories.

AN OVERVIEW OF ECCO

The first thing to note when comparing ECCO to the ESTC is that not every document in ECCO directly corresponds to a single record in the ESTC. This is because multi-volume works and periodicals have only a single record in the ESTC, whereas all volumes are recorded separately in ECCO. In this article, we mainly operate on the less granular level of the ESTC. In total, our version of ECCO1 contains 153,924 documents, which correspond to 136,164 distinct ESTC records, and our version of ECCO2 contains 52,689 documents, corresponding to 48,222 ESTC records. With respect to time, apart from individual outliers, the ECCO documents range from 1701 to 1800. Our version of the ESTC contains a total of 344,759 records for this time period. Thus, our first conclusion is that ECCO by no means covers all of the eighteenth-century publications included in the ESTC. Overall, ECCO1 appears to cover 40 percent of the ESTC for the period 1701–1800, rising to 54 percent with the addition of ECCO2.

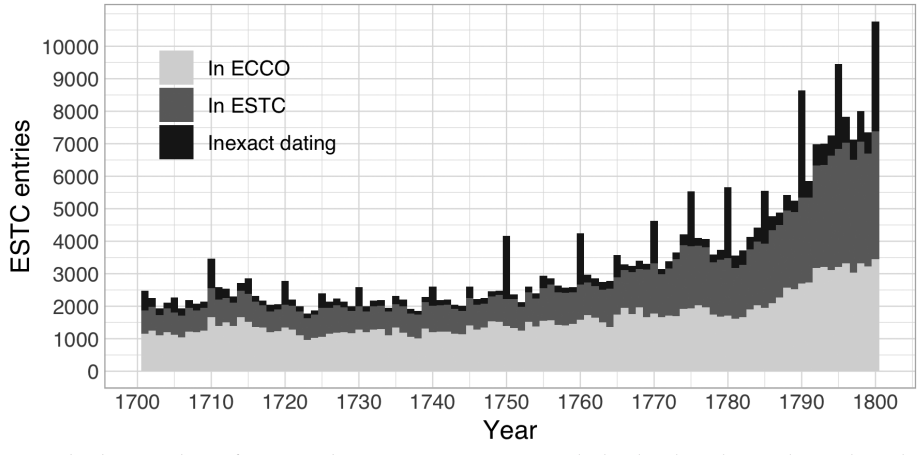

Fig. 1 Absolute numbers of ESTC and ECCO entries over time. The bar height indicates the total number of works in the ESTC per year. Visualized separately are records whose dates in reality constitute a time-span (on top), records not found in ECCO (in between) and records included in ECCO (at the bottom).

It has often been suggested that the overall statistical development of printing in early modern Britain was one of exponential growth.[39] With reference to the ESTC and ECCO, Figure 1 shows the overall distribution of ESTC entries over the eighteenth century, as well as the proportion of those entries appearing in ECCO. Count spikes every even decade, in some mid-decade years as well as in 1750 and 1800, appear due to uncertainties in dating the historic materials. In both ECCO as well as the ESTC, when an exact date of publication is not known, catalogers still often record approximate or probable dates. Such dates are usually meant to be interpreted as inexact, covering a range of possible dates around the stated value. Although ECCO does not include structured information on dating accuracy, in the ESTC, such dates are marked with e.g., a *ca.,* or a *(?)*. With the help of this ESTC metadata we are able to identify which of the entries have inexact dates. In total, such records constitute approximately 10 percent of ECCO and 14 percent of overall ESTC records for the eighteenth century. Interpreting such dates as exact may strongly affect any analyses intended to trace temporal phenomena in the corpus. 32 percent of ECCO records and a full 47 percent of ESTC records for 1750 are marked as approximate, the implied dating for most of them probably being any time in the eighteenth century. We removed the inexactly dated records from our analysis when the focus was on temporal changes, otherwise we have retained them.

As mentioned, with the addition of ECCO2, 54 percent of the records contained in the ESTC are reproduced in ECCO. However, the question still remains whether all these records are distributed equally on all axes of interest, or whether there are imbalances in the representation. It is hard to see from the absolute counts depicted in Figure 1 whether or not the proportional share of ESTC entries making their way into ECCO changes from year to year. For this reason, we report the coverage in terms of percentages for each year in all subsequent temporal analyses. Such views do reveal the change in representation. In particular, despite the rapid growth over the course of the eighteenth century in the total number of print products captured in the ESTC, particularly after 1780, these products do not make their way equally to the ECCO.

To exemplify both absolute and proportional amounts, let us consider the years 1710 and 1790. The ESTC lists 2,559 unambiguously dated records for 1710, of which 1,663 or 65 percent are in ECCO. For the year 1790, on the other hand, the number of unambiguously dated ESTC records (5,341) is more than double that of 1710. However, of these, only 51 percent (2,699) are in ECCO. Thus, although ECCO contains much more material from the end of the century in absolute terms, in terms of the proportion of what would be available in libraries, it contains less. Both of these findings could lead to faulty analytical conclusions, but whether or not they do would depend on the research questions. For example, any analysis taking ECCO to represent eighteenth-century language as a whole would be biased towards more modern vocabulary based on the rise in absolute publication numbers. Comparing the first half of the century and the latter half is problematic because the proportions of different types of material excluded from ECCO changed over time, as discussed in the "Books and pamphlets" section below.

BOOKS AND PAMPHLETS

Whether the print product was a shorter pamphlet or a longer book turned out to be a determining factor in whether it was in ECCO. It is impossible to give a precise definition of an eighteenth-century pamphlet or book that covers every case. According to Halasz, "neither 'pamphlet' nor 'book' is a generic category, but rather, an indicator of object form that slides easily into commodity designation (and dismissal)."[40] Our interest is not to arrive at a general definition of a pamphlet as such, but rather to study smaller and larger forms of publishing under the denominators of pamphlet- and book-sized print products. This is in sync with Halasz's observation that "'pamphlet' functions as a floating signifier in the heterogeneity that characterizes the opportunities made available by print."[41] It is also understandable that there is no clear distinction between pamphlets and small books. Pamphlets (as well as books) are often categorized intuitively.

For our purposes, following experimentation with different ways of categorizing the data, we eventually ended at a cut-off point of a maximum 32 pages for a pamphlet-sized document, and a minimum 128 pages for a book, regardless of the leaf size.[42] Setting separate cut-off points for pamphlets and books ensures high homogeneity within these categories, while leaving in between a fuzzy category comprising works of between 33 and 127 pages in length. Defined in this way, pamphlets constitute 50 percent of the ESTC for the whole of the eighteenth century, whereas books and in-between works account for 25 percent each. These percentages remain remarkably constant through time, even when the total volume of publications increases drastically after 1780 (Figure 1).

In what follows, we have enacted our analyses on all three of the abovementioned categories. However, we consistently present the results only for pamphlets and books, leaving out the in-between category. We do this in the interest of presentational clarity given that, in terms of behavior, we found the in-between category to consistently follow the book category, but with more variability caused by its fuzzier nature.

In terms of the representation of pamphlets and books in ECCO, Figure 2 depicts our first insights. As already stated, in this and the following figures, the main Y-axis has been switched from reporting the absolute counts of works shown in Figure 1 to reporting how large a percentage of the works reported in the ESTC for each year is in ECCO. However, because it is also important to not lose sight of the differences in absolute counts, they are represented in this and following figures as circle sizes where possible. From Figure 2, it is clear that book-length print products have a much higher chance of making their way into ECCO than pamphlets. However, it is also evident that, although the proportional representation of books is relatively equal over time (with a modest linear decline from around 75 to 65 percent in coverage between 1700 and 1800), the coverage of pamphlets is more varied, and drops significantly between 1765 and 1775, from around 50 to around 30 percent. Bearing in mind the rapid growth in print publications at the end of the century (shown here as the sizes of the circles, but more easily seen in Figure 1), this means that although the number of books in ECCO increases in tandem with the increase in the ESTC at the end of the century, the number of pamphlets in ECCO does not increase to the same extent. Thus, proportionally

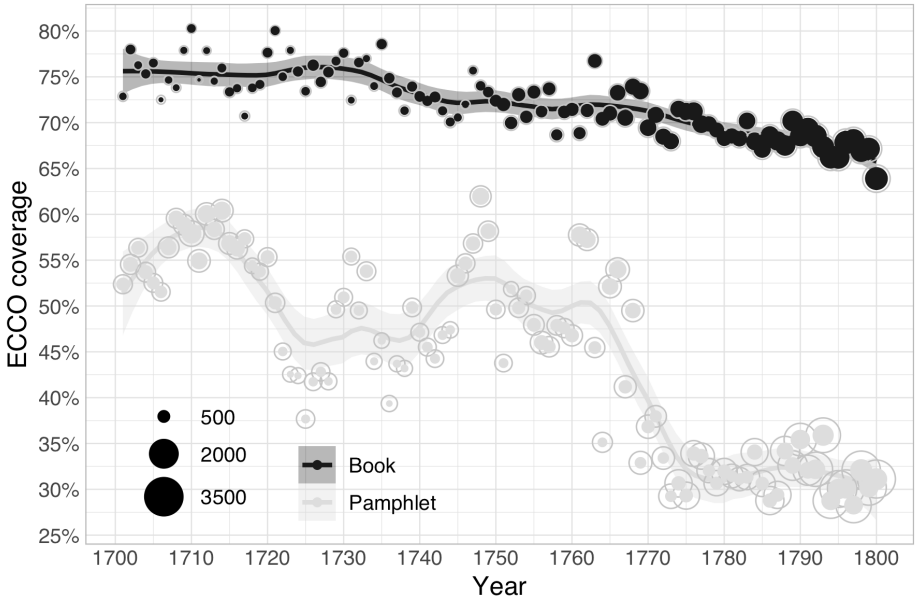

Fig. 2 ECCO coverage per publication format during the eighteenth century. Yearly coverages are indicated as circles, and the loess-smoothed trend highlights the main patterns. Circle size depicts the number of publications. The filled portions of the circles depict the number of publications in ECCO, and the outlines show the numbers in the ESTC.

far fewer of the pamphlets produced in the late eighteenth century have made their way into ECCO. As a consequence, without corrective measures, any ECCO-based analyses of the printing of pamphlets or the overall nature of printing in the eighteenth-century Anglophone world cannot be accurate because of the large proportion of missing pamphlets.

Delving further into the uneven coverage of pamphlets, we sought to find out whether particular types of pamphlets were being excluded. It is difficult to assess whether ECCO covers different subjects equally. The ESTC uses the Library of Congress Subject Headings (LCSH), which is very fine-grained: there are 9,425 distinct headings referenced in our data, most of which appear only a few times. Nevertheless, through a series of mappings we were able to match these to Gale's categories, although with limitations.[43] First, 48 percent of the eighteenth-century records in the ESTC lack keywords altogether, thereby prohibiting any mapping. Second, our mapping accuracy is sufficient only for some of Gale's categories. The Fine Arts, History and Geography, and General Reference categories had to be dropped completely due to unreliable mappings, raising the number of unmapped entries to 49 percent. Further, we noticed in our evaluation that the Law and Social Sciences categories were often confused with each other, so we decided to merge them in our analysis.

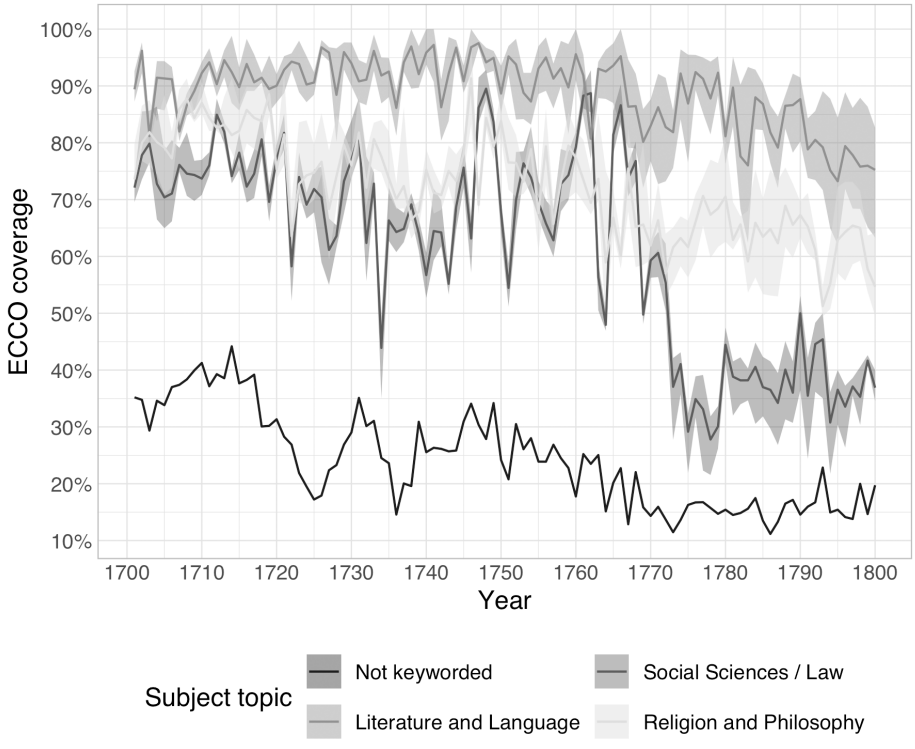

Fig. 3 The representation of selected subject topics for pamphlets. The shaded areas depict minimum and maximum percentages of the ECCO coverage of different pamphlet types obtained using different methods. The line depicts the mean percentage.[44]

Operating within these limitations, we noticed that ECCO covers literary pamphlets to a higher degree than other types of pamphlets (Figure 3). Of the other reliably identified categories, Religion and Philosophy together with Medicine, Science, and Technology shared almost equal representation throughout this period. Hence, for the sake of clarity, we only show the former category in Figure 3. A further effect revealed in Figure 3 is that having been keyworded in the ESTC clearly correlates with a higher probability of the pamphlet being in ECCO. In fact, this effect was also evident for books, albeit the difference in percentage points being "only" up to 30 instead of the 60 for pamphlets. Finally, after 1770 there was a particularly steep drop in the coverage of pamphlets categorized as social sciences or law, this being the only category to correlate with the large drop in overall pamphlet coverage at that time. It seems from a random sample of the excluded pamphlets in this category that many are governmental acts and proclamations, lending credence to our assumption that it is precisely such documents that were no longer included in ECCO after that point in time.

We also identified two further imbalances in ECCO with respect to publication formats. First, within the category of person authors (i.e., not organizations), ECCO is highly imbalanced towards including pamphlets by people who also authored at least one book-length publication: the mean chance of such people getting their pamphlets into ECCO is around 70 percent, compared with only around 40 percent among those who only authored pamphlets (regardless of how many).

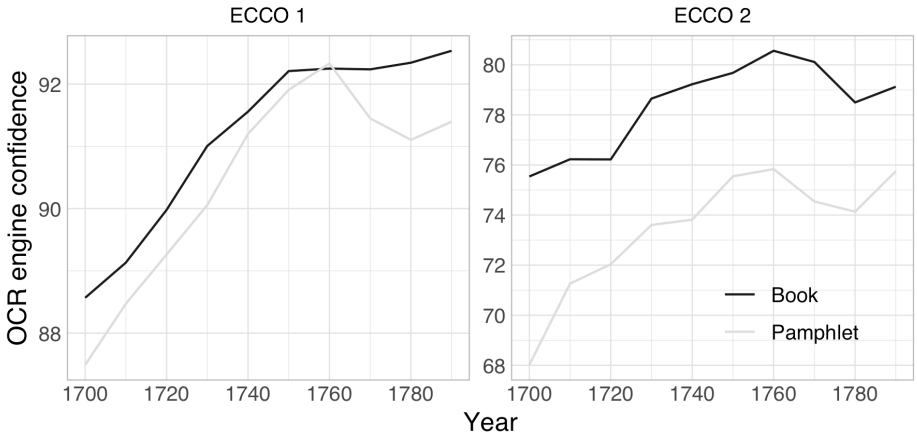

Fig. 4 Mean OCR engine confidence for pamphlets and books in ECCO. Note that the confidence scores reported by the engine used for ECCO1 are not directly comparable to the scores reported by the engine used for ECCO2.

Our final observation related to publication format is not about total inclusion or exclusion, but rather concerns imbalances in the OCR quality of the publications that were included. Here, as measured by the confidence levels given by Gale's OCR engines and visualized in Figure 4, the mean OCR quality does not differ between pamphlets and books in ECCO1.[45] But in ECCO2, which uses a different OCR engine, pamphlets consistently are of a significantly lower OCR quality than books.[46] Otherwise, in both collections, OCR quality increases in publications from the latter part of the century, with a particularly pronounced effect in ECCO1. This means that in addition to being affected by the representational biases and gaps mentioned above, keyword searches are also more likely to find matches in books as opposed to pamphlets, as well as from the latter part of the century as opposed to the earlier.

PLACES OF PUBLICATION

Fifty-nine percent of the eighteenth-century records in the ESTC, and 64 percent in ECCO, are publications that were printed in London. London's dominance aside, there is clear regional variance in the datasets. For example, German-language publications in Pennsylvania played an important role in shaping the cultural landscape for at least a generation.[47] These kinds of observations are important and it would be a great mistake to ignore publishing activity outside London in any quantitative study of the eighteenth century.[48] Hence, it is interesting to see whether places outside London receive equal consideration. As we show below, some places of publication are particularly poorly represented in the data, and some of the cities cannot be studied in any reliable manner based on ECCO data.

Country	ESTC	ECCO	Pamphlets Covered	Books Covered
England	233,526	134,946 (58%)	43%	77%
USA	40,686	10,088 (25%)	18%	43%
Scotland	33,879	17,366 (51%)	40%	75%
Ireland	25,000	16,650 (67%)	59%	78%
France	2,527	1,398 (55%)	40%	70%
Canada	995	35 (4%)	1%	13%
Others	4,539	2,158 (28%)	16%	33%
Total	**341,152**	**182,641 (54%)**	**40%**	**74%**

Table 1. Publications by country in the ESTC and ECCO. The coverage refers to the proportion of works listed in the ESTC that are also covered in ECCO. Books and pamphlets differ systematically in terms of coverage.

As Table 1 shows, on the country level, almost 75 percent of all the titles included in ECCO were published in England (compared to 68% for the ESTC). The reason for this higher figure in ECCO is that the coverage of American print products is very low—only 25 percent for the United States and four percent for Canada. The number of documents missing from the United States in ECCO compared with the ESTC exceeds 30,000, which is more than the number of missing documents from all the other countries combined, excluding England.

It should be borne in mind, of course, that as we have shown, many official and legal documents are purposely excluded from ECCO. Hence, in the cases of England and Scotland in particular, it is better to look at the ECCO coverage of books to understand the value of ECCO as a representative research object. This coverage is fairly similar in England, Scotland, and Ireland, ranging between 75 and 78 percent. The colonies, on the other hand, are more heavily underrepresented in terms of book coverage, which is a little over 40 percent in the United States and a miserable 13 percent in Canada. In the case of pamphlet-sized documents, the United States coverage drops to below 20 percent and Canadian coverage almost to zero, whereas Irish coverage climbs to almost 60 percent.

In terms of temporal differences in spatial coverages (Figure 5), Irish coverage is consistently over 60 percent, reaching an average closer to 70 percent towards the end of the eighteenth century. Coverage in England falls steadily over the later part of the century, but still remains above 50 percent at the end. Scottish coverage in ECCO is below 50 percent for most of the early eighteenth century, but approaches that of England later.

In terms of imbalance, North American representation in the data stands out. United States coverage is consistently under 40 percent, and even lower in the later eighteenth century. This could be partly attributable to the fact that the geographical spread of material largely depends on which institutions provided scans for ECCO. Although American libraries were part of the microfilm project underlying ECCO, it is clear that they remain heavily underrepresented in the dataset.

As pointed out above, a more fine-grained city-level analysis shows that eighteenth-century publishing was centered in but not confined to London. There

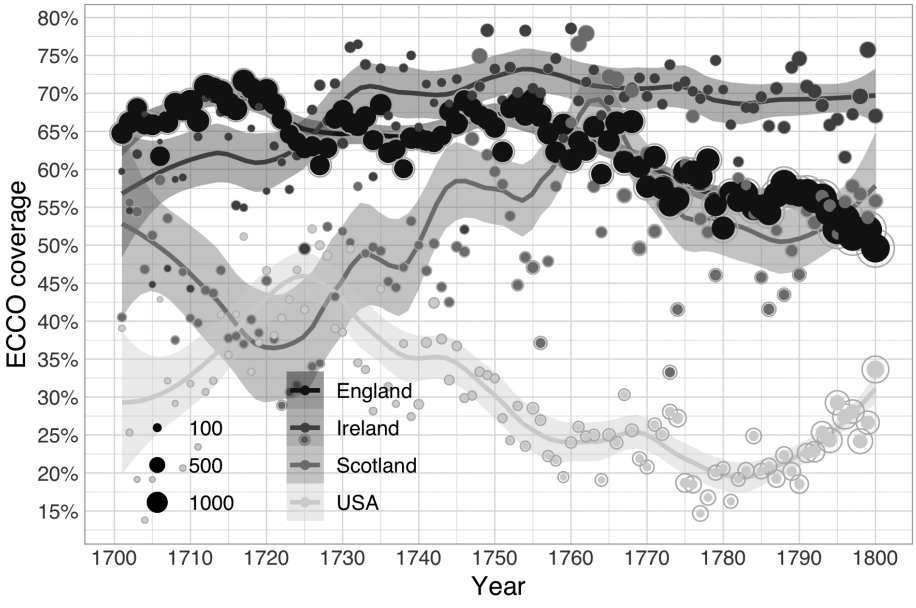

Fig. 5 Temporal variations in publishing coverage in different countries. Annual coverage is indicated by circles, and the loess-smoothed line highlights the main trend. The filled portions of the circles indicate the number of publications in ECCO, and the outlines show the corresponding ESTC numbers.

was a clear trend for London books to be targeted at other centers in England, and distributed and sold elsewhere as well, especially in the second half of the century.[49] Publishing in provincial towns is closely associated with reprints.[50] With regard just to the ESTC, James Raven has conducted a basic analysis of places of publication.[51] Aside from the reprints, official governmental status is reflected in the total publication volumes in particular places. According to Brown and McDougall, for example, as much as 75 percent of the total annual printing output from Edinburgh concerned official governmental and other legal documents.[52]

City	ESTC	ECCO	Pamphlets Covered	Books Covered
London	202,174	117,681 (58%)	43%	77%
Edinburgh	26,657	13,710 (51%)	41%	78%
Dublin	23,722	15,809 (67%)	59%	78%
Philadelphia PA	10,167	2,241 (22%)	13%	40%
Boston MA	9,834	3,987 (41%)	32%	49%
Glasgow	5,015	2,621 (52%)	34%	69%
New York NY	4,537	808 (18%)	9%	41%
Oxford	3,612	2,381 (66%)	58%	77%
Others	55,434	23,403 (35%)	22%	31%
Total	**341,152**	**182,641 (54%)**	**40%**	**74%**

Table 2. Publications by city in the ESTC and ECCO.

In terms of our materials, the overall publication numbers per city (Table 2) reflect the numbers per country (Table 1). The reason for this correlation is that London dominates English printing, Dublin dominates Irish printing, and Edinburgh dominates Scottish printing; hence, London, Dublin, and Edinburgh follow the same pattern as England, Ireland, and Scotland respectively in terms of coverage (Table 1). However, what the city perspective adds is the inclusion of Oxford, Glasgow, and three American cities. The same imbalance with respect to the United States is also evident with respect to these American cities. Of these, Boston has the best coverage in ECCO, although its book coverage is still less than 50 percent on average. Philadelphia boasts a larger number of records in the ESTC than Boston, but it has very low ECCO coverage. ECCO's coverage of print products originating from eighteenth-century New York is also dismal. Oxford, on the other hand, fares almost as well as Dublin when it comes to ECCO coverage, even with respect to pamphlets. Glasgow, known as a reprint city, is covered fairly well.[53]

EDITIONS AND REPRINTS

Gale calls ECCO a "critical tool" and a collection that "contains every significant English-language and foreign-language title printed in the United Kingdom between the years 1701 and 1800."[54] The idea that ECCO is mainly about items limited to first and significant editions of each title is a legacy of the original microfilm project led by Alston in the 1980s at the British Library.[55] In this section we will show that the reality is different.

It has been well established in book history that reprints dominated the eighteenth-century printing business.[56] Dublin reprints of London books comprised a large segment of the eighteenth-century book market.[57] The relationship between the provinces and London turned out to be difficult as provincial printing started to accelerate in the mid-eighteenth century. Scotland's ascendance as a printing center in the 1750s was based on both legitimate and pirated reprints.[58] The 1710 copyright act was not enforced in the United States either.[59] In this context, it is striking that users of the ECCO dataset are left under the impression that it mainly includes first and further editions that have significant additions in comparison to the first printing.[60] In light of our recent harmonization of the ESTC edition field, we are now able to evaluate, first, whether reprints or first/singular editions of works have a higher chance of being represented in ECCO, and, second, what proportion of the publications consists of reprints.[61]

To obtain a data-driven picture of the coverage of different edition types, we would ideally want to divide the data into singular publications (meaning work only printed once), first editions, further editions with significant additions, and near-identical reprints. Unfortunately, our ESTC data does not allow us to differentiate between reprints and significant editions. To get past this problem, we derived a paired set of proxy variables that still allow us to approach these questions. As one end point, we keep all editions of a work separate. In the following, we refer to these simply as editions. As the other end point, we aggregate all editions of a work published in the same year, seeing whether at least one of those editions has made the cut to ECCO. In the following, we refer to these aggregates simply as works. Further, we differentiate between first-year and later editions and works.

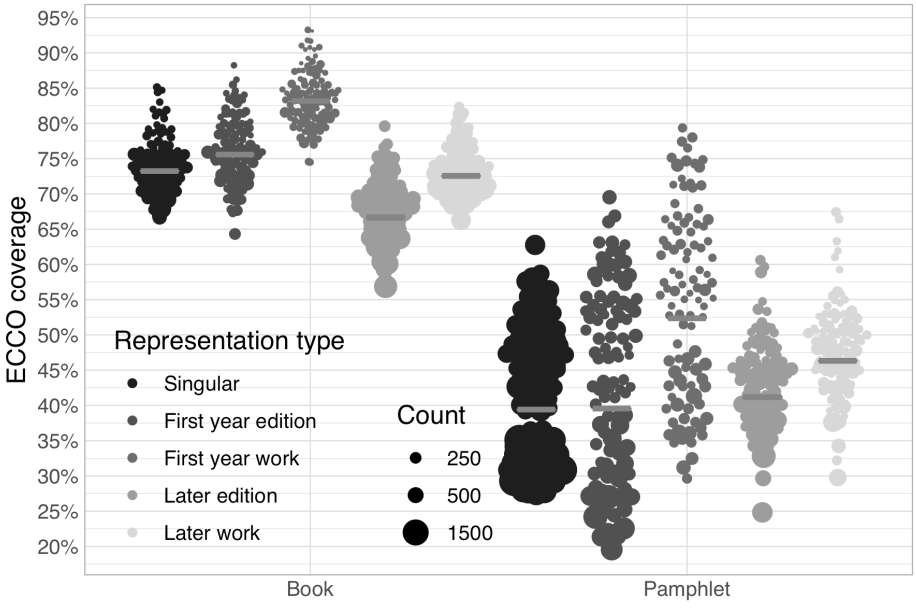

Fig. 6 Representation in ECCO by publication format and representation type. Each point corresponds to a single year, the size denoting the number of publications of that type for that year. The horizontal bars denote the median coverage for the given type of representation.

The intuition here is that if there is a tendency in ECCO to only include first and significant editions, we are likely to see 1) first-year editions and works be better represented than later-year editions and works and 2) a significantly higher representation on the work level as compared to the edition level for both first-year as well as later data points. As Figure 6 shows, with regard to pamphlets, at the edition level there are no differences in median coverage between the singular, first-year, and later editions. On the work level, however, at least one of the first-year editions is much more likely to be included than works of just a single edition. Similarly, with regard to later editions of the pamphlets, at least one version per year is slightly more likely to be included than when each edition is counted separately. This may point toward the hypothesis of ECCO including only first and significant editions of pamphlets. On the other hand, it may also be explained by ECCO including only "important" pamphlets, i.e. pamphlets of the type liable to get reprinted in the first place.

In the case of books, differences appear already on the edition level, with later annual editions being statistically less likely to be included in ECCO than either first-year or singular editions. As in the case of pamphlets, aggregation of the editions onto the work level by year also increases the chance of at least one edition making it into ECCO. With regard to later annual reprints, however, this boost only raises their representation up to par with the singular-edition works.

Unfortunately, we cannot ascertain from this analysis whether the work-level effects reflect a clear agenda aimed at prioritizing first editions, or whether they derive purely from the natural accumulation of probability when aggregating multiple editions. What is clear, however, is that on the work level one is significantly more likely to find keyword matches in works that were more popular in the eighteenth century, for example, thus leading to their having multiple editions

in ECCO. Flipping the question however, given the small difference between work-level and edition-level coverage for both first-year and later data points, we can already deduce that ECCO must contain a significant amount of duplication due to the inclusion of reprints, thereby shattering any naive thinking otherwise and the advertising of ECCO being to any relevant sense duplicate-free.

Turning our analysis in precisely this direction, we find that of the 183,758 ESTC records found in ECCO, 67,989 are identified by our pipeline as further editions of works that ECCO already contains. Therefore, approximately 37 percent of the publications in ECCO are potential duplicates. Delving more deeply, we also see that the amount of duplication varies significantly by publication type: of the book-length publications included in ECCO, a full 51 percent are further editions of works it already contains, whereas only 21 percent of the pamphlets are duplicates. Given the differences in page count between books and pamphlets, the duplication of content on the word level is significantly higher than the 37 percent of work duplication. Indeed, according to our evaluation, a massive 51 percent of the content of ECCO measured in words is duplicated.

Such scale of duplication will have an impact on any large-scale text mining. For example, the Cambridge Concept Lab, which is one of the groups that has advanced furthest in the use of ECCO for text mining, has worked under the assumption that the significance of the reprint phenomenon is limited, and would not meaningfully affect the results of large-scale text mining.[62] In our view, this is a shortcut, and researchers should not close their eyes to data variation and imbalance. We suggest that, instead, the inclusion or exclusion of reprints in studies that rely on ECCO must be decided on the basis of the analysis target. If the idea is to understand what was available to the general public at a particular time, the inclusion of reprints can be useful. But if the aim of a study is to understand how many later authors were affected by a novel idea, then reprints should be excluded. Similarly, if one is interested in finding out how language evolves, one needs to be aware of the magnitude of reprints and how findings might be thereby affected.

Given the high numbers of reprints across the categories, it would be of interest, particularly to people wishing to use ECCO for diachronic studies, to know the approximate age of the reprints. According to our analysis, half of them are reprints of works less than 20 years old. However, there is also a long tail, as well as differences between publication formats. Of the pamphlets, 18 percent are more than 50 years old, and the percentage of book reprints of works from over 50 years prior is even higher at 25 percent. Furthermore, about 10 percent of all the reprints in ECCO are of works originally published more than 100 years preceding the reprint. Once again, given the differences in page count, it should be borne in mind that in terms of text mass, old books contribute significantly to the representational imbalance.

AUTHORS

The question of the emergence of the "modern" author in the eighteenth century is complex, and imprints do not always give a full picture of the author/writer of each ESTC record.[63] Anonymous publishing (i.e., withholding the name of the writer of the work on the title page) was extremely common until the nineteenth century.[64] Of the editions in ECCO, for example, 27 percent (51,519 publications)

are anonymous or by otherwise unknown authors. Moreover, publishers often controlled how the author was represented on the imprint, and many authors, especially if they were female, wrote under a pen name.[65] Our particular focus is on known authors. Using the author's name as a unit of analysis enables us to differentiate groups of texts from each other in the ESTC and ECCO, to establish relationships among the texts, and, perhaps, to even identify discourses.[66] Even if such theoretical concepts of an author tend to be highly ahistoric, the opportunity to analyze the development of authorship statistically from imprint information will fundamentally enhance understanding of eighteenth-century Britain. Thus, the question of imbalances in ECCO with respect to different authors recorded in ESTC imprints is relevant to any user of the database.

One identifiable imbalance in ECCO with regard to authors is the tendency of prominent authors to be disproportionately more likely to be included in ECCO than lesser-known ones. So, if we are under the impression that ECCO as such represents eighteenth-century public discourse and treat it naively, we keep on putting unnecessary emphasis on those authors whose editions are overrepresented in the collection in comparison to authors in general. This imbalance can lead to cumulative effects with close resemblance to the so-called *Matthew Effect*, or the "rich get richer" phenomenon.[67] In the context of this article, we hypothesized that more popular authors might receive inordinately more frequent coverage in ECCO and hence be more likely to accumulate visibility and citations of their work than authors who are less popular. "Popular authors" here are simply writers with the most editions recorded in the ESTC. We used this to investigate how ECCO coverage varied by author popularity. More specifically, we applied binomial regression to estimate the probability of works being included in ECCO as a function of author popularity. The probabilistic analysis proved to be a valuable tool in that it allowed us directly to compare coverages among authors who had very different numbers of works and editions, while at the same time taking into account the remarkable variations in the ECCO coverage between different authors. Indeed, the analysis showed a significantly increasing trend in the probability of ECCO coverage among the more popular authors, thus confirming our working hypothesis. Of those authors who only had a single published edition in the ESTC, only about 50 percent had their works included in ECCO, whereas those with more than ten editions had, on average, over 65 percent of them included in ECCO. Hence, the more popular authors appear to have a disproportionately higher coverage in ECCO.

However, upon closer inspection we found that this imbalance seemed not to be distinct from the imbalances we identified earlier. First, the majority of authors with only a single edition are authors of pamphlets, which as already noted have, in general, poorer coverage than books. Second, the authors who did publish more than one work tended to publish both books and pamphlets; as reported above, the probability of such authors having their pamphlets included was higher than among authors who only published pamphlets. Of these two effects, the former seems an unrelated feature of the material. But in the case of the latter effect, selections in the scanning process may have been guided by whether or not the pamphlets came from well-known authors. Or, what is perhaps more likely, selections may have been influenced by the works that were physically located in the libraries in the same collections as the book-length works of the same authors.

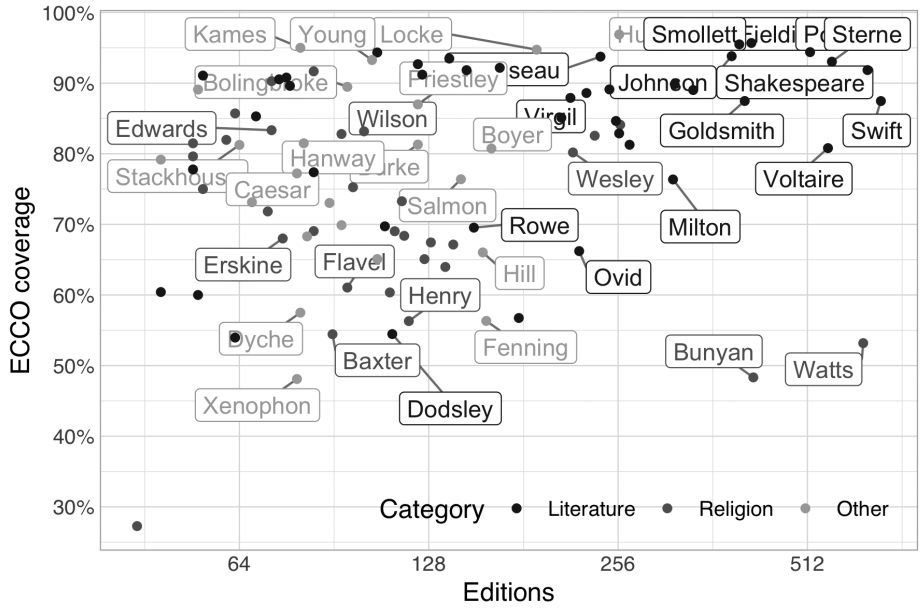

Fig. 7 The most prominent authors per subject topic. Author prominence ranking has been calculated for all authors in the ESTC. It takes into consideration number of works and editions, years in print, and geographical spread. Author categories have been chosen based on the most common genre of publications for authors with works in multiple genres. This figure captures 95 authors in total (41 literature, 31 religion, and 23 other); not all have been labeled.

In general, works by the most popular British authors are well covered in ECCO (Figure 7). ECCO has an average coverage of 84 percent for the 1,000 authors with the most editions; these authors cover 40 percent of all the editions (54,935) included in ECCO by known authors. There is a clear difference between *edition coverage* and *work coverage*: the former refers to the presence of editions in ECCO, and the latter means that at least one edition of a particular work is included in the collection. Thus, the coverage of works is bound to be much higher than that of editions. The editions of several authors are covered comprehensively if not fully in ECCO. For example, if we look at the most popular authors of literature in the figure above, we may note that as a group they are even better represented than religious authors. Also, a total of 22 works of John Locke (1632–1704) appear in the ESTC, of which only three are missing from ECCO. Also, many other famous authors such as George Berkeley (1685–1753), Elizabeth Rowe (1674–1737), and William Robertson (1721–1793) have all of their works that are listed in the ESTC also represented in ECCO.

In contrast to previously mentioned names, the coverage in ECCO of some authors with an equal claim to fame is much thinner. The influential American preacher Cotton Mather (1663–1728) is among the authors with the most individual works listed in the ESTC, but his coverage in ECCO is low (65%). By way of comparison, the work coverage in ECCO of Daniel Defoe (1661–1731), whose oeuvre includes several pamphlets as well as popular prose, is 86 percent. Although Methodist preacher George Whitefield (1714–1770) and immensely productive hymn writer Isaac Watts (1674–1748) are among the most popular authors based on records in the ESTC, their coverage in ECCO is less than 70 percent. Whitefield's case is particularly interesting because the other leading Methodist author

of the eighteenth century, John Wesley (1703–1791), is well covered in ECCO. According to Isabel Rivers, Watts was the "most important figure in the development of eighteenth-century hymnody."[68] The fact that many of his works are not included in ECCO shows that this prominence does not necessarily come through in the dataset. According to our author statistics, one particularly poorly covered group of authors is American political writers: less than 50 percent of works by American politicians such as Thomas Pownall (1722–1805), George Washington (1732–1799), John Dickinson (1732–1808), Thomas Jefferson (1743–1826), Alexander Hamilton (1757–1804), and Robert Goodloe Harper (1765–1825) are covered in ECCO. One consequence of this imbalance is that whatever works of these poorly represented groups of authors are found may be mistaken for comprehensive sources of information if users are under the impression that ECCO contains all the relevant data.

John Wesley is the author with the most editions of all of his works (907) listed in ECCO. Not all religious authors are underrepresented, however: 43 of the 44 editions that Presbyterian minister and historian Edmund Calamy (1671–1732) had printed in the eighteenth century are represented in ECCO, for example. The fact that there is a considerable amount of religious material missing from ECCO further points to the possibly mistaken significance of these religious authors whose coverage exceeds that of other religious authors. It also underscores the elitist nature of ECCO—it does not provide a direct path to the public discourse of eighteenth-century Britain. This imbalance is even more pronounced when we consider what is thoroughly covered in ECCO. In the case of Locke, for example, 161 documents are listed in the ESTC with him as an author, of which only 24 are missing from ECCO. Thus, it could be said that Locke's presence in ECCO is comprehensive (86% of all the editions recorded in the ESTC). The average edition coverage for the 1,000 authors with the most editions included in ECCO is 72 percent. We found many similar cases of authors with a high output in terms of editions (close to 100 or more) and extensive coverage (close to 85% or more). Among these are authors one might expect to be included, such as Alexander Pope (1688–1744), Henry Fielding (1707–1754), Samuel Johnson (1709–1784), David Hume (1711–1776), Laurence Sterne (1713–1768) and Tobias Smollett (1721–1771).

FEMALE AUTHORS

There is a possibility that even a larger number of female authors are missing from the ESTC records than what we have estimated based on the information that we have been able to collect. It is therefore possible that, from the outset, there was an overall imbalance with respect to gender even before we started to compare the ESTC and ECCO records. This is a good reason to analyze female authors separately.[69] According to Bell, there has been a tendency to underestimate the amount of women's publications in the eighteenth century.[70] We have attempted also to include voices identified as female (such as publications by "A Lady") in the overall analysis of female output, but we have not included them in the analysis of individual female authors because we do not know how many different authors lie behind a pseudonym such as "A Lady." Calculated in this manner, according to Grundy: "The numbers [of female publications] per decade, having risen fairly steadily during the first half of the century, took off around 1760, and thereafter

virtually doubled every decade."[71] What is of relevance here is how the ECCO and ESTC numbers correspond, and if this gendered imbalance is further amplified in ECCO. Our recently published statistical analysis supports such a conclusion, although it seems that the effect of gender is smaller than that of publication format.[72] The remaining effect could also be largely—albeit not fully—attributed to the tendency among female authors to publish shorter documents, which are less likely than book-sized objects to be included in ECCO. We already know that most famous eighteenth-century authors are overrepresented in ECCO, and that the works of many other authors are missing. With respect to female authors, a few individuals stand out with respect to the total number of works included in ECCO: Eliza Haywood (1745–1833), Hannah More (1745–1833), Susanna Centlivre (1667?–1723), and Elizabeth Singer Rowe (1674–1737).

We cannot use the average coverage of the editions, works, books, and pamphlets of female authors when we aim to understand the overall variation in the data. Some authors are covered well, but if we consider all 3,200 or so female authors, the averages are pushed down because of the low output and coverage of some of them. This does not apply only to female authors: the result is the same if we include all authors in our calculations, because many produced only a few works. At the same time, the editions of some female authors are covered almost in full: the twenty-nine editions of English novelist/poet Clara Reeve (1729–1807) listed in ECCO is equivalent to a 94 percent coverage of all her editions in the ESTC.

POPULAR WORKS

If we turn our attention from authors to individual works (Figure 8), major differences begin to arise in how works are represented in ECCO. Of the works by a known author with the highest edition count, John and Charles Wesley's *Hymns* has over 140 editions listed. Other works with high edition counts include Alexander Pope's *Essay on Man* (113), John Milton's (1608–1674) *Paradise lost* (118), and Defoe's *Robinson Crusoe* (116). Coverage of the editions of all of these works is above 50 percent compared to the ESTC.

Signs of imbalance emerge when we look at how some of the most extensively reprinted eighteenth-century religious works are represented in ECCO. *Pilgrim's progress*, *Whole duty of man*, *Divine songs*, *Practice of pietie*, and Watts's *Hymns* were very popular at the time, but their coverage is low. This is also true of William Lily's (1468–1522) *Grammar*. Since Gale states that ECCO does not contain ephemera, it is understandable that the inclusion of sale catalogs to ECCO is incidental, but religious works are different. Although almost 100 editions of *Robinson Crusoe* are missing from ECCO, 116 are still included. There is a stark contrast here with John Bunyan's (1628–1688) *Pilgrim's progress*: 34 editions are listed in ECCO, but over 200 have been left out, marking a clear imbalance in ECCO against the book. Of the 194 editions of Benjamin West's (1730–1813) *Almanacs* recorded in the ESTC, only 11 percent make their way to ECCO. The inclusion of almanacs in ECCO is sporadic, but in West's case it did not help that his works were printed in the United States.

Switching our focus from works with the highest edition count to works that are most heavily represented in ECCO reveals 10 works with 50 editions or

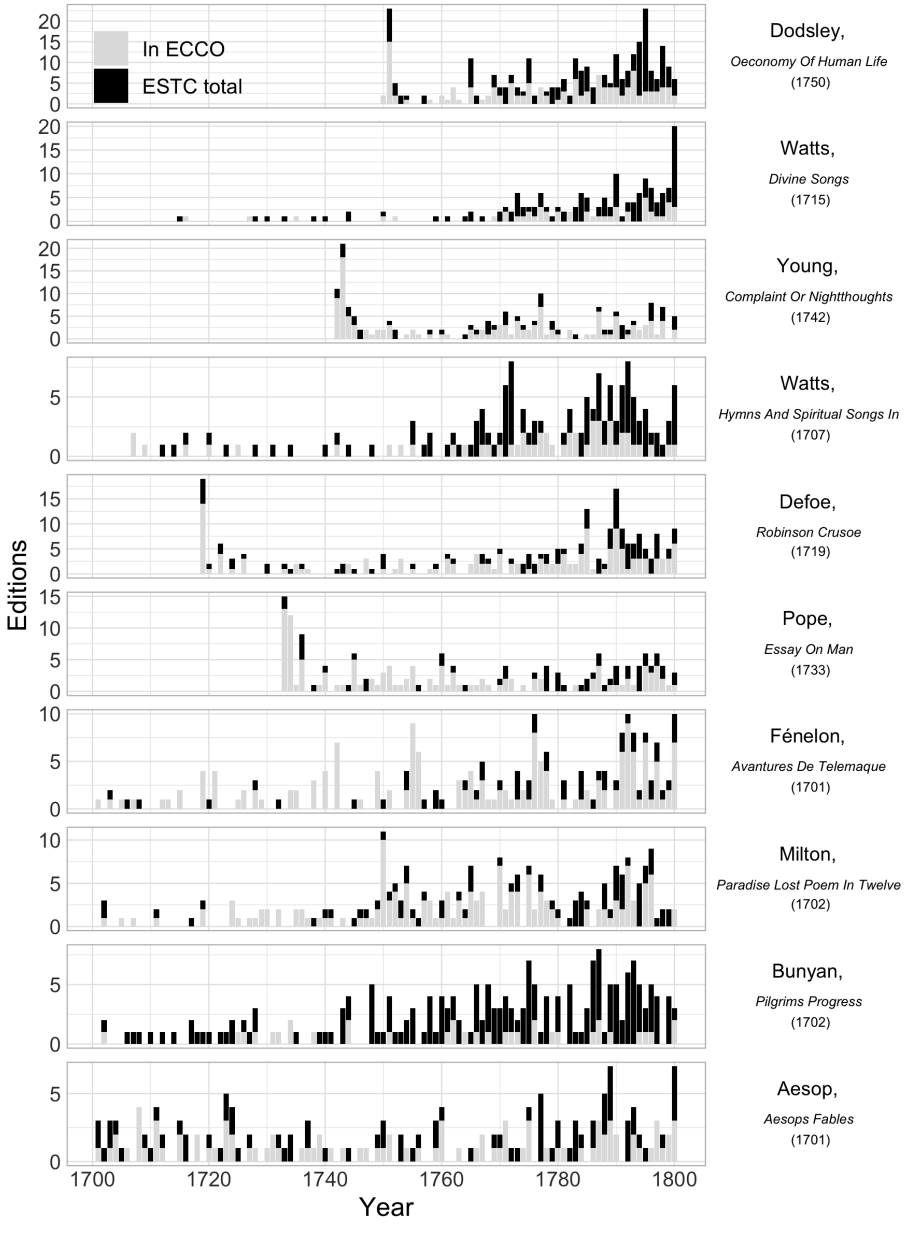

Fig. 8 Top works in ECCO. All works in the ESTC have been calculated with a work prominence ranking that takes into consideration number of editions, years in print, and geographical spread. Year after the title denotes the first recorded publication in ECCO.

more and a coverage of 80 percent or higher. Only six of the 56 editions of the *Abridgement* of Locke's *Essay concerning Human understanding* are missing. Other frequently reprinted works with a high coverage include Edward Young's (1683–1765) *Universal passion* and Samuel Butler's (1613–1680) *Hudibras*. A full 78 editions of Jonathan Swift's (1667–1745) *Miscellanies* are included in ECCO, covering almost 80 percent of the editions listed in the ESTC. Other works worth mentioning with both a high number of editions and extensive coverage in ECCO

include *Letters of Junius*, Paul de Rapin's (1661–1725) *History of England*, Miguel de Cervantes's (1547–1616) *Don Quixote*, August von Kotzebue's (1761–1819) *Spanier in Peru*, and Alain René Le Sage's (1668–1747) *Gil blas*. Mary Collyer's (c. 1716–1763) translation of *Death of Abel in five books* is also comprehensively covered with 29 editions out of 32 included in ECCO. The contrast is stark in the case of Isaac Watts's *New England Primer,* which was reprinted more often than any of these works but only has 10 percent coverage in ECCO. Thomas Dilworth's (–1780) vastly popular *A new guide to the English tongue* has almost sunk into oblivion in ECCO, with only 16 percent of its eighty-one eighteenth-century printings covered. The main reason for this is that, just as in the case of Noah Webster's (1758–1843) *Grammatical institute of English language*, a substantial number of these editions were printed in the United States and are not included in the microfilm collection that forms the basis of ECCO.

We have already established that the coverage of pamphlets in ECCO is lower than that of books, especially if they are more frequently reprinted. Nevertheless, most editions of some particular pamphlet-length print products are included. This mainly concerns short works of literature: the above-mentioned Edward Young's (1683–1765) *Universal passion* was often reprinted in pamphlet length, as were certain political and religious treatises such as Charles Lucas's (1713–1771) *To free citizens and free holders of the city of Dublin* and John Wesley's *Nature design and general rules of united societies*. Each of these works have a coverage of more than 80 percent in ECCO.

CONCLUSION

This article has demonstrated that there are considerable imbalances in ECCO that need to be taken into consideration in all uses of its data. Pointing out gaps is not to undermine the overall value of the database. Its coverage of different works, especially those that are book-sized, is high; a researcher is therefore likely to find at least the most common sources in it. Moreover, several editions of works written by the most popular authors are generally available, which is, after all, what most users (especially students, who are one of Gale's major target audiences) are after. Yet at the same time, the imbalances we have revealed have substantial implications for anyone using the collection. There are also many overlapping sources of imbalance that cumulate, and individual works are best compared against their overall background (i.e., coverage of a particular work compared to the general coverage of other similar works). We admit that systematic comparison of each individual case may be difficult, although our work provides some guidance and methodology.[73] All ECCO users will have to consider for themselves the potential effects of the varying imbalance and representativeness in the ESTC and ECCO on their own work.

The main types of imbalance we identified concern the temporal aspects of the data, publication formats, spatial aspects, editions and reprints, and authors (including questions of gender and the popularity of individual works). With respect to publication formats, the longer the work, the more likely it is to be included in ECCO. The differences between the early and the later eighteenth century need to be taken into consideration in temporal analyses in particular. There is a general

drop in coverage after the 1770s, which is largely attributable to a drop in pamphlet coverage: there was only a slight decrease in books at the same time. There are also imbalances with regard to the subject topics in the pamphlets. Religion, for example, consistently features less than literature, whereas governmental acts and proclamations account for most of the difference in coverage before and after the 1770s.

With regard to reprinting, the representational imbalances in ECCO are considerable and should be taken into consideration whenever the database is used. The claim that ECCO focuses mainly on first printings and editions with substantial changes simply does not hold water when examined more closely.

On the regional level, North American publications—especially those from the United States—are heavily underrepresented in ECCO, while Ireland is exceptionally well represented. In terms of publication formats, pamphlets are missing way beyond the ephemera that were not supposed to be included in ECCO in the first place. This overall imbalance is further amplified in that some genres are overrepresented in pamphlets. The main imbalances with regard to authors concern genres: popular authors of literature are overrepresented, whereas the representation of religious authors is more sporadic. Moreover, even the most popular female authors are underrepresented, in spite of the fact that ECCO2 was supposed to correct this.

How seriously different users ought to consider the anatomy of ECCO depends, of course, on their needs and research questions. As we have shown, there are enormous variations in author representation in ECCO that can lead us unwittingly to emphasize such authors that are overrepresented in the data. Not only does it matter that we cannot find authors that we ought to in the context of our searches, but perhaps an even stickier problem is that the overrepresented ones pop up everywhere. A comprehensive evaluation of user needs is beyond the scope of this article. We do make the point, however, that every single user should be aware of the main points that we have uncovered to facilitate use of the database and the related analytical tools. As we implied in the introduction, no dataset will ever be able to reflect historical reality as such. It is up to users to decide which imbalances are problematic, and which they are free to ignore in a given study. Naturally, ECCO will keep evolving through the introduction of ECCO3 and other advances in the future. Perhaps also our efforts to "reconnect" the ESTC and ECCO will improve the quality of the dataset. In any case, a better understanding of the anatomy of ECCO and its relevant imbalances will go a long way to improve research using the Eighteenth Century Collections Online.

NOTES

1. Open access funded by Helsinki University Library.

2. On the history of ECCO, see especially Stephen Gregg, *A History of Eighteenth Century Collections Online; or, Old Books and Digital Publishing* (Cambridge: Cambridge Univ. Press, 2020). On its relevance to digital humanities, see Emily Friedman, "Afterword: Novel Knowledge, or Cleansing Dirty Data: Toward Open-Source Histories of the Novel," in Ileana Baird, ed., *Data Visualization in Enlightenment Literature and Culture* (London: Palgrave Macmillan, 2021), 351–70. A substantially narrower paper targeting corpus linguists in particular based on the research leading to this article has been published earlier in Mikko Tolonen, Eetu Mäkelä, Ali Ijaz, and Leo Lahti, "Corpus Linguistics and Eighteenth Century Collections Online (ECCO)," in *Research in Corpus Linguistics* 9 (2021): 19–34.

3. For early discussions on ECCO, read especially the *Eighteenth-Century Intelligencer*; e.g. Sayre Greenfield, "EC/ASECS Presidential Address, October 2006, Gettysburg: 'ECCO-Locating the Eighteenth Century,'" *Eighteenth-Century Intelligencer*, 21, no. 1 (Jan. 2007): 1–9; Robert Hume's "The ECCO Revolution," *Eighteenth-Century Intelligencer* 21, no. 1 (Jan. 2007): 9–17; and Corey E. Andrews, "ECCO and the Future of Eighteenth-Century Studies," *Eighteenth-Century Intelligencer*, 22, no. 2 (2008): 8–13.

4. For the full timeline for the history of ECCO, see Gregg, *A History of Eighteenth Century Collections Online*, v–vi. The ESTC comprises three earlier catalogs: The Short-Title Catalogue (STC, which covers the period 1473–1640), Wing (covering 1641–1700), and The Eighteenth Century Short Title Catalogue. For more details, see David Vander Meulen, "ESTC as Foundational and Always Developing," *Age of Johnson* 21 (2011): 263–82; David McKitterick, "'Not in STC': Opportunities and Challenges in the ESTC," *The Library* 7th ser. 6, no. 2 (2005): 178–94.

5. Robin Alston, "The History of the ESTC," *Age of Johnson* 15 (2004): 269–329; Robin Alston and Mervyn Jannetta, *Bibliography, Machine-Readable Cataloguing and the ESTC* (London: The British Library, 1978).

6. Welly Kinley, "Digital ECCOs of the Eighteenth Century," *eContent* (November 2003).

7. Cassidy Holahan, "Rummaging in the Dark: ECCO as Opaque Digital Archive," *Eighteenth-Century Studies* 54, no. 4 (2021): 803–26.

8. Eileen Clancy, et al., "Beyond Citation - Eighteenth Century Collections Online (ECCO)," Beyond citation (2014), https://www.beyondcitation.org/database/eighteenth-century-collections-online-ecco/.

9. Gale, "Eighteenth Century Collections Online," *ECCO roll fold* (brochure) 2016, https://www.gale.com/binaries/content/assets/gale-us-en/primary-sources/eighteenth-century-collections-online/ecco-roll-fold-2016-web.pdf.

10. For a discussion about collections and modeling in digital humanities, see Katherine Bode, "Why You Can't Model Away Bias," *Modern Language Quarterly* 81 no. 1 (2020): 95–124.

11. McKitterick, "'Not in STC'": 178.

12. In what follows, all our calculations are based on XML data dumps of ECCO1 and ECCO2 obtained from Gale in 2015 through the Helsinki University Library, in accordance with Gale's updated text-mining policy that allows researchers of a subscribing institution access to the content outside of Gale's user interface. All comparisons to the ESTC are against our offline version, graciously provided to us by the British Library in March 2016 and updated in September 2020.

13. James E. May sees the dependence of ECCO to the ESTC as a problem due to the mistakes in the ESTC, see May, "Some Problems in ECCO (and ESTC)" *Eighteenth-Century Intelligencer* 23, no. 1 (2009): 20–30. In contrast, we see bibliographic metadata as evolving materials that we harmonize ourselves, aware of the initial shortcomings in the data.

14. Clancy, et al., "Beyond Citation," 2014.

15. For use of ECCO in teaching in the 2000s, see forum of "Papers on Teaching with ECCO and EEBO," *Eighteenth-Century Intelligencer* 23, no. 3 (2009): 2–29.

16. Stephen Gregg, "Digital Humanities and Archives @ ASECS 2012," *Manicule* (blog), April 5, 2012, https://shgregg.com/2012/04/05/digital-humanities-and-archives-asecs-2012–3/.

17. Kelly Centrelli, "ECCO on JISC and Contextual Word Searches," *The Long Eighteenth* (blog), April 10, 2012, https://long18th.wordpress.com/2012/04/10/ecco-on-jisc-and-contextual-word-searches/.

18. See Patrick Spedding, "'The New Machine': Discovering the Limits of ECCO," *Eighteenth-Century Studies* 44, no. 4 (2011): 437–53; Iain Gadd, "The Use and Misuse of Early English Books Online," *Literature Compass* 6, no. 3 (2009): 680–92; and Sayre Greenfield, "ECCO OCR Trouble-shooting," *Early Modern Online Bibliography* (blog), 2010: https://earlymodernonlinebib.wordpress.com/ecco-ocr-troubleshooting-by-sayre-greenfield/.

19. Eetu Mäkelä, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi, and Terttu Nevalainen, "Wrangling with Non-Standard Data," in Sanita Reinsone, Inguna Skadiņa, Anda Baklāne, and Jānis Daugavietis, eds., *Proceedings of the Digital Humanities in the Nordic Countries 2020, CEUR Workshop Proceedings* 261 (2020): 81–96.

20. In the sphere of statistics, problems of dataset imbalance such as these are often referred to as "dataset bias." In this specialist use, the term "bias" appears as a neutral technical term, used to denote any aspect of a research design that may invalidate the neutrality of the inference process.

21. In her survey, Holahan identified that already twenty percent of the rising number of people who cited ECCO as part of their research used it in such a manner. Holahan, "Rummaging in the Dark": 811.

22. cf. Peter de Bolla, Ewan Jones, Paul Nulty, Gabriel Recchia, and John Regan, "The Idea of Liberty, 1600–1800: A Distributional Concept Analysis," *Journal of the History of Ideas* 81 (2020): 381–406; and a working group called Arguing with the Digital History, "Digital History and Argument," *Roy Rosenzweig Center for History and New Media* (white paper), November 13, 2017, https://rrchnm.org/argument-white-paper/.

23. Linguistic DNA was an Arts & Humanities Research Council (AHRC) funded project that functioned as a collaboration between English historical linguists at the universities of Sheffield, Glasgow, and Sussex, The Digital Humanities Institute and the Historical Thesaurus of the Oxford English Dictionary. https://www.linguisticdna.org/ We are currently working on an additional article analyzing the representativeness of EEBO and EEBO-TCP.

24. For more information about COMHIS, see: https://www2.helsinki.fi/en/researchgroups/computational-history.

25. Paddy Bullard, "Digital Humanities and Electronic Resources in the Long Eighteenth Century," *Literature Compass* 10 (2013): 749.

26. Iiro Rastas, Yann Ryan, Iiro Tiihonen, Mohammadreza Qaraei, Liina Repo, Rohit Babbar, Eetu Mäkelä, Mikko Tolonen, and Filip Ginter, "Explainable Publication Year Prediction of Eighteenth Century Texts with the BERT Model," in Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin, eds., *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change* (Stroudsburg: The Association for Computational Linguistics, 2022): 68–77.

27. Seth Cayley, "Digitization for the Masses: Taking Users Beyond Simple Searching in Nineteenth-Century Collections Online," *Journal of Victorian Culture* 22, no. 2 (2017): 249.

28. Gale, "Digital Scholar Lab," *Gale Primary Sources* (website), https://www.gale.com/intl/primary-sources/digital-scholar-lab and "Discover the Possibilities. Explore Primary Sources Through a New Lens," *Gale Digital Scholar Lab* (brochure), October 2021.

29. It is not the objective of this essay to explain why music printing, for example, is not sufficiently featured in the ESTC. Rather, we offer a comparative analysis of the available data. For historical contextualization, see Gregg, *A History of Eighteenth Century Collections Online*.

30. G. Thomas Tanselle, "Bibliography and Science," *Studies in Bibliography* 27 (1974): 55–90.

31. Alston, "History of the ESTC": 318–19.

32. Mikko Tolonen, Mark Hill, Ali Ijaz, Ville Vaara, and Leo Lahti, "Examining the Early Modern Canon: The English Short Title Catalogue and Large-Scale Patterns of Cultural Production," in Baird, ed., *Data Visualization in Enlightenment Literature and Culture*, 63–119; Leo Lahti, Niko Ilomäki, and Mikko Tolonen, "A Quantitative Study of History in the English Short Title Catalogue (ESTC) 1470–1800," *LIBER Quarterly* 25, no. 2 (2015): 87–116; Leo Lahti, Jani Marjanen, Hege Roivainen, and Mikko Tolonen, "Bibliographic Data Science and the History of the Book (c. 1500–1800)," *Cataloging & Classification Quarterly* 57, no. 11 (2019): 5–23.

33. Hugh Amory, "Pseudodoxis Bibliographica, or When Is a Book Not a Book? When It's a Record," in Lotte Hellinga, ed., *The Scholar and the Database, CERL Papers 2* (London: Consortium of European Research Libraries, 2001), 7.

34. Vander Meulen, "ESTC as Foundational and Always Developing": 263; Leo Lahti, Eetu Mäkelä, and Mikko Tolonen, "Quantifying Bias and Uncertainty in Historical Data Collections with Probabilistic Programming," in F. Karsdorp, B. McGillivray, A. Nerghes, and M. Wevers, eds., *Proceedings of the Workshop on Computational Humanities Research CHR2020,* CEUR workshop proceedings 2723 (2020): 280–89; Mikko Tolonen, Leo Lahti, Jani Marjanen, and Hege Roivainen, "A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828," *Historical Methods* 52 (2018): 57–78; Lahti et al., "Bibliographic Data Science and the History of the Book": 5–23; Lahti et al., "A quantitative Study of History in the English Short Title Catalogue (ESTC) 1470–1800": 87–116. The Thomason Tracts constitute an example of the possible over-representation of pamphlets in the ESTC, see Iiro Tiihonen, "From Explosion to Implosion. A Quantitative Analysis of the English Civil War Print," MA thesis, Univ. of Helsinki, Faculty of Arts (2020).

35. Stephen W. Brown and Warren McDougall, "Introduction," in Stephen W. Brown and Warren McDougall, eds., *The Edinburgh History of the Book in Scotland*, vol. 2. *Enlightenment and Expansion. 1707–1800* (Edinburgh: Edinburgh Univ. Press), 14–18.

36. Alexandra Hill, *Lost Books and Printing in London, 1557–1640: An Analysis of the Stationers' Company Register* (London: Brill, 2018).

37. Amory, "Pseudodoxis Bibliographica," 8.

38. Robert K. Merton, "The Matthew Effect in Science," *Science* 159 (3810), (1968): 56–63.

39. Eltjo Buringh and Jan Luiten van Zanden, "Charting the 'Rise of the West': Manuscripts and Printed Books in Europe a Long-Term Perspective from the Sixth through Eighteenth Centuries," *The Journal of Economic History* 69, no. 2 (2009): 409–45; Joerg Baten and Jan Luiten van Zanden, "Book Production and the Onset of Modern Economic Growth," *Journal of Economic Growth* 13, no. 3 (2008): 217–35; Michael F. Suarez, "Towards a Bibliometric Analysis of the Surviving Record 1701–1800," in Michael F. Suarez and Michael L. Turner, eds., *The Cambridge History of the Book in Britain,* vol. 5,. *1695–1830* (Cambridge: Cambridge Univ. Press, 2009), 37–65; Tolonen et al., "Examining the Early Modern Canon," 63–90.

40. Alexandra Halasz, *The Marketplace of Print: Pamphlets and the Public Sphere in Early Modern England* (Cambridge: Cambridge Univ. Press, 1997), 3. See also Joad Raymond, *Pamphlets and Pamphleteering in Early Modern Britain* (Cambridge: Cambridge Univ. Press, 2003); and Jason Peacey, "Pamphlets," in Joad Raymond, ed., *The Oxford History of Popular Print Culture, vol. 1, CheapPrint in Britain and Ireland to 1660* (Oxford: Oxford Univ. Press, 2011), 453–71.

41. Halasz, *The Marketplace of Print*, 186.

42. See Mikko Tolonen, Eetu Mäkelä, and Leo Lahti, "Supplementary Information to The Anatomy of Eighteenth Century Collections Online (ECCO) article," *Supplement to 'Anatomy of ECCO,'* June 2022. https://doi.org/10.5281/zenodo.6683914.

43. For the details, see Tolonen, Mäkelä, and Lahti, "Supplementary Information to The Anatomy of Eighteenth Century Collections Online (ECCO) article."

44. For the details, see Tolonen, Mäkelä, and Lahti, "Supplementary Information to The Anatomy of Eighteenth Century Collections Online (ECCO) article."

45. For the details on why this can be trusted, see Tolonen, Mäkelä and Lahti, "Supplementary Information to The Anatomy of Eighteenth Century Collections Online (ECCO) article".

46. Note that the confidence scores reported by the two engines are not directly comparable between each other. Thus, from Figure 4, one cannot draw, for example, the conclusion that the overall OCR quality in ECCO1 would be better than in ECCO2. Instead, the opposite is very much likely the case, as the engine used to OCR ECCO2 is much newer. Evaluating this would need an external evaluation set that covers both corpora, which we did not have available at the time of writing.

47. Gregg A. Roeber, "German and Dutch Books and Printing," in Hugh Armory and David D. Hall, eds., *A History of the Book in America, vol. 1, The Colonial Book in the Atlantic World* (Chapel Hill: The Univ. of North Carolina Press, 2007), 298–313. Our analysis of different countries (including North America) is naturally based on the ESTC because ECCO was built off it. Other union catalogs are available, however, and comparisons (in the North American case to the North American Imprints

Program [NAIP], for example) may be beneficial in the future. In the use of NAIP for statistical purposes, see Hugh Amory, "A Note on Statistics," in Armory and Hall, eds., *A History of the Book in America*,vol. 1, 504–18; and David Hall and Russell Martin, "A Note on Popular and Durable uthors and Titles," in Armory and Hall, eds., *A History of the Book in America, vol. 1,* 519–21.

48. John Feather, "The British Book Market 1600–1800," in Simon Eliot and Jonathan Rose, eds., *A Companion to the History of the Book*, 2nd edition (London: Wiley & Sons, 2020), 414–20; and Michael F. Suarez, "Book History from escriptive Bibliographies," in Leslie Howsam ed., *The Cambridge Companion to the History of the Book*, (Cambridge: Cambridge University Press, 2014), 199–218.

49. Feather, "The British Book Market 1600–1800," 414–20.

50. Catherine Armstrong and John Hinks, eds., *Printing Places: Locations of Book Production & Distribution Since 1500* (New Castle, DE: Oak Knoll Press and the British Library, 2005).

51. James Raven, "Investing in Books: the Supremacy of the Booksellers," in *The Business of Books: Booksellers and the English Book Trade* (New Haven: Yale Univ. Press, 2007), 149–53.

52. Brown and McDougall, "Introduction," 15.

53. Michael Moss, "Glasgow," in Brown and McDougall, eds., *The Edinburgh History of the Book in Scotland, vol.* 2, 156–65.

54. Gale, "Eighteenth Century Collections Online," *ECCO roll fold* (brochure) 2016, https://www.gale.com/binaries/content/assets/gale-us-en/primary-sources/eighteenth-century-collections-online/ecco-roll-fold-2016-web.pdf; and Gale, "Eighteenth Century Collections Online. Eighteenth-Century Britain: Published," *Gale Primary Sources* (website), https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online.

55. Gregg, *A History of Eighteenth Century Collections Online*, 22–24.

56. Thomas F. Bonnell, "Reprint Trade," in Suarez and Turner, eds., *The Cambridge History of the Book in Britain, vol. 5,* 699–703 and John Feather, *A History of British Publishing* (London: Routledge, 1998).

57. John Feather, "The Publishers and the Pirates. British Copyright Law in Theory and Practice, 1710–1775," *Publishing History* 22 (1987): 17; and Charles Benson, "The Irish Trade," in Suarez and Turner, eds., *The Cambridge History of the Book in Britain, vol. 5,* 370.

58. Feather, "The Publishers and the Pirates," 17–19; Warren McDougall, "Copyright and Scottishness," in Brown and McDougall, eds., *The Edinburgh History of the Book in Scotland, vol.* 2, 23–39.

59. James N. Green, "The British Book in North America," in Suarez and Turner, eds., *The Cambridge History of the Book in Britain, vol. 5,* 548–49; and James N. Green, "English Books and Printing in the Age of Franklin," in Armory and Hall, eds., *A History of the Book in America, vol.* 1, 248–98.

60. Gregg, *A History of Eighteenth Century Collections Online*, 23.

61. Ali Ijaz, Leo Lahti, Iiro Tiihonen, and Mikko Tolonen, "Analytical Determination of Editions from Bibliographic etadata," in *Proceedings of the Research Data And Humanities (RDHUM) 2019 Conference: Data, Methods And Tools* (Studia Humaniora Ouluensia 17, 2019).

62. Peter de Bolla, Ewan Jones, Paul Nulty, Gabriel Recchia, and John Regan, "The Conceptual Foundations of the Modern Idea of Government in the British Eighteenth Century: A Distributional Concept Analysis," *International Journal for History, Culture and Modernity* 7 (2019): 619–52.

63. Martha Woodmansee, "The Genius and the Copyright. Economic and Legal Conditions of the Emergence of the 'Author,'" *Eighteenth-Century Studies* 17, no. 4 (1984): 425–48.

64. Robert Griffin, "Anonymity and Authorship," *New Literary History* 30, no. 4 (1999): 877–95.

65. Dustin Griffin, "The Rise of the Professional Author?" in Suarez and Turner, eds., *The Cambridge History of the Book in Britain, vol. 5,* 132–45; David Scott Kastan, "The Emergence of the Author," in David Loewenstein and Janet Mueller, eds., *The Cambridge History of Early Modern Literature*

(Cambridge: Cambridge Univ. Press, 2002), 108–16; Stephen B. Dobranski, *Readers and Authorship in Early Modern England* (Cambridge: Cambridge Univ. Press, 2005); Maureen Bell, "Women Writing and Women Written," in J. Barnard and D. McKenzie, eds., *The Cambridge History of the Book in Britain, vol. 4, 1557–1695* (Cambridge: Cambridge Univ. Press, 2002), 431–52.

66. Michel Foucault, "What is an Author?," in *Language, Counter-Memory, Practice: Selected Essays and Interviews*, translated by D.F. Bouchard and S. Simon (Cornell: Cornell Univ. Press, 1977), 123.

67. About the Matthew Effect on a general level, see Robert K. Merton, "The Matthew Effect in Science," *Science* 159 (1968): 56–63. In the context of historical research, attention has been drawn to the limitations of online sources and an overemphasis on what has been digitized, cf. Ian Milligan, "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010," *The Canadian Historical Review* 94, no. 4 (2013): 540–69. Our point about the Matthew Effect is somewhat different, intended to show that author popularity in the eighteenth century already played a major role in the creation of representational imbalances.

68. Isabel Rivers, "Religion and Literature," in John Richetti, ed., *Cambridge History of English Literature, 1660–1780* (Cambridge: Cambridge Univ. Press, 2005), 468.

69. For a discussion on the use of short-title catalogs when estimating numbers of female authors, see Maureen Bell, "Women and the Production of Texts: the Impact of the History of the Book," in John Hinks and Victoria Gardner, eds., *The Book Trade in Early Modern England: Practices, Perceptions, Connections* (New Castle, DE: Oak Knoll Press and the British Library, 2014), 107–22.

70. Bell, "Women Writing and Women Written," 432–33. See also Catherine Gallagher, *Nobody's Story: the Vanishing Acts of Women Writers in the Marketplace, 1670–1820* (Oxford: Oxford Univ. Press, 1995).

71. Isobel Grundy, "Women and Print: Readers, Writers and the Market," in Suarez and Turner, eds., *The Cambridge History of the Book in Britain, vol. 5,* 146–47. See also Judith Phillips Stanton, "Statistical Profile of Women Writing in English from 1660 to 1800," in Frederick M. Keener and Susan E. Lorsch, eds., *Eighteenth-Century Women and the Arts* (New York: Greenwood Press, 1988), 247–54.

72. Lahti et al., "Quantifying Bias and Uncertainty in Historical Data Collections," 284–86.

73. See also Tolonen, Mäkelä, and Lahti, "Supplementary Information to The Anatomy of Eighteenth Century Collections Online (ECCO) article."