LIBER QUARTERLY
THE JOURNAL OF THE ASSOCIATION OF EUROPEAN RESEARCH LIBRARIES

Vol. 25, no. 2 (2015) 87–116 | ISSN: 1435-5205 | e-ISSN: 2213-056X

A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470–1800[1]

Leo Lahti

Laboratory of Microbiology, Wageningen University, The Netherlands
leo.lahti@iki.fi

Niko Ilomäki

Department of Computer Science, University of Helsinki, Finland
niko.ilomaki@cs.helsinki.fi

Mikko Tolonen

Department of Modern Languages, University of Helsinki, Finland
mikko.tolonen@helsinki.fi

Abstract

This article analyses publication trends in the field of history in early modern Britain and North America in 1470–1800, based on English Short-Title Catalogue (ESTC) data.[2] Its major contribution is to demonstrate the potential of digitized library catalogues as an essential scholastic tool and part of reproducible research. We also introduce a novel way of quantitatively analysing a particular trend in book production, namely the publishing of works in the field of history. The study is also our first experimental analysis of paper consumption in early modern book production, and demonstrates in practice the importance of open-science principles for library and information science. Three main research questions are addressed: 1) who wrote history; 2) where history was published; and 3) how publishing changed over time in early modern Britain and North America. In terms

of our main findings we demonstrate that the average book size of history publications decreased over time, and that the octavo-sized book was the rising star in the eighteenth century, which is a true indication of expanding audiences. The article also compares different aspects of the most popular writers on history, such as Edmund Burke and David Hume. Although focusing on history, these findings may reflect more widespread publishing trends in the early modern era. We show how some of the key questions in this field can be addressed through the quantitative analysis of large-scale bibliographic data collections.[3]

Key Words: history publishing; short-title catalogue

1. Introduction

Library catalogues are a prominent information source for anyone studying the history of publishing and the associated social change. The standardized large-scale nature of these data collections, cataloguing up to millions of documents, calls for an automated quantitative framework that could be used to shed light on the development of book production in the early modern world, for example.[4] Library catalogues have conventionally been seen as a tool for finding a particular item in the library system.[5] We demonstrate how these catalogues can be used not only as research tools, but also as research objects. We have integrated parts of the English Short Title Catalogue (ESTC) from the British Library with statistical algorithms to establish a data-analytical ecosystem via which to analyse changes in book production in early modern Britain and North America during the hand press era (1470–1800).[6]

We propose using such open data-analytical ecosystems, libraries of a kind, to supplement and explore the full contents of digital data resources with state-of-the-art data-analysis techniques.[7] This would notably complement conventional database-query interfaces such as the ESTC, EEBO and ECCO, which allow specific searches but not a full exploration of the database contents. Such arbitrary restrictions on data availability place severe limitations on how the data can be used, and create a significant bottleneck for data-driven research. Moreover, even when data collections are made available, the lack of statistical tools specifically designed for such analysis constitutes another practical research obstacle. The ecosystems we propose combine

research data and custom algorithms to enable flexible and deep quantitative analysis of the full data collections. These systems can be implemented in open-source software libraries that are developed as part of the research process and provide customised tools for analysing data collections of interest. Here we demonstrate the benefits of such an approach in a practical case study on knowledge production in the field of history during 1470–1800.

2. Library catalogues as a research resource

According to Peter Stallybrass (2004), researchers should be turning to librarians to understand knowledge production (see also Kraus, 1986). It is well known that union catalogues such as the ESTC can be used as a research resource for statistical analysis, although they were not originally designed for such a purpose (Suarez, 2009).[8] We propose a new way of carrying out such research and demonstrate how library catalogues can provide a valuable data resource for historical studies. We show how the analysis can be performed within a comprehensive quantitative framework, following the best open-science practices of transparency, reproducibility and code sharing.

Our analysis covers documents in the ESTC catalogue that include the word 'history' in any of the catalogued subject fields covering the years 1473–1800. This includes 50,766 entries among the 466,000 documents catalogued in the ESTC (~10%). We do not aim at an exhaustive or objective classification of history as a genre (or genres) as represented in the catalogue (and we acknowledge the limitation that it does not include all documents published in Britain and North America from 1470 to 1800).[9] Nevertheless, anyone interested in studying David Hume's *History of England*, for example, will now have convenient, adaptive tools for quantitative analysis based on bibliographic catalogues that will help considerably to situate this work in the context of general knowledge production. The generic principles of openness and automation we promote here could be extended later to the mining of full-text databases to gain further insights into the historical evolution of terms, concepts and research topics.[10]

Although the quantitative approach to the history of the early modern book has been recommended on several occasions (Weedon, 2007; see also the pioneering work of Bell and Barnard, 1992, 1998), it has not been implemented to

the extent that is possible.[11] Some scholars constructing quantitative analyses of book history have been rather sceptical about their own approach.[12] One reason for this is that, as many scholars indicate, library catalogues do not represent a stable database, but are constantly being changed and updated, leaving the results from large-scale studies vulnerable to change in the database contents (see, in particular, Karian, (2011); Raven (2014) was also critical). This is precisely why there is a need for automated open workflows such as the one implemented in this research, whereby results can easily and automatically be updated when new versions of the data arrive. Thus, in the case of the ESTC for example, should plans to formulate an improved ESTC21 catalogue be realised one day, our tools could easily be applied to the new version.[13] Large-scale analysis of general trends complements the analysis of specific documents, authors or publication periods by setting them in the wider context of overall knowledge production. Moreover, the analysis of large-scale statistical trends in knowledge production can be expected to be robust against specific database updates. We believe that our methods will significantly expand the use of a quantitative framework for qualitative research. At the same time, it is important to acknowledge the limitations of such approach; the overall information content of the catalogue and supporting information sources that can be linked to the study (May, 1984) sets limits on what the analysis can provide.[14]

To our knowledge, this is the first organized plan to move towards a comprehensive transparent quantitative framework within which to study the history of the book. An article on the bibliometric analysis of surviving records published as recently as 2009 (Suarez, 2009), for example, does not propose any solutions to the problems of transparency, automation and reproducibility that we have resolved here.

3. Open-data analytical ecosystems as quantitative research tools

The starting point of our analysis is the library catalogue, which is further extracted, transformed, and supplemented with supporting information, and subjected to rigorous quantitative analysis. The analysis is based on custom data analysis algorithms that are implemented as part of the research project. The combination of data and algorithms constitutes an 'ecosystem' that can

be further refined and extended by updating the research data or source code. The algorithms provide generic research tools that can be potentially used beyond the scope of the original study.

The open-data analytical ecosystems we propose here ideally include: (i) the full database contents in an open, machine-readable format, provided by the institution that holds the data; (ii) supporting data sources, preferably from open source repositories; and (iii) well-documented open source algorithms to extract relevant information from the data and transform it into the final statistical summaries and visualizations in a fully automated, reproducible and transparent manner. The ESTC represents the main data source of interest in our case, further supported by external data sources such as publicly available name-gender mappings, geographical coordinate databases, custom lists of author pseudonyms, and other supplementary sources that support the interpretation, as we demonstrate in more detail below. The algorithms and the complete analysis workflow are being made available via the ESTC R package on the Github social coding platform (https://github.com/ropengov/estc) that facilitates further community contributions and feedback.

A central element of our work is that we make the full algorithmic details openly available for anyone to use, verify and improve further.[15] The source code provides a detailed description of all the steps, from the data to the final quantitative results, tables and figures. Whereas the source code implementing a specific analysis is typically newly created and customized in each research project, many algorithms for specific analysis tasks can be readily borrowed from existing open source libraries. This leaves the researchers more time to focus on the new research questions, and thus makes the research more efficient. At the same time, our original contributions within this project have been publicly shared from the very beginning. Ideally, other scholars and the general public will be able to use the algorithms to study related research questions, or as a starting point to develop further tools and find new uses in other contexts.

We have implemented the work in the R statistical programming environment (https://www.r-project.org/), which is already widely used in other fields of science. This enables seamless integration of the data sets with state-of-the-art data-analysis techniques, and allows researchers to build their own research tools by combining existing standard algorithms with a custom source code, specifically designed for the given research project. In contrast

to commercial software suites such as Matlab, SAS or SPSS, our approach is fully open source and provides dedicated tools for large-scale analyses of library catalogues. Unlike standard query interfaces such as ESTC, EEBO or ECCO, which provide only limited access to the database and do not allow large-scale data mining, our approach takes advantage of the complete data contents of the library catalogue.

4. Who wrote history?

It is interesting to ponder on the question of who wrote history, and on whether a quantitative analysis of publication volumes and numbers of imprints would support the common understanding of the most famous historians who published in the English language.[16] A key challenge in this analysis is that the same author may be listed under multiple variants of the name. To overcome this we implemented parsers that remove special characters and recognize first and last names based on large background lists from public databases and manually prepared supplementary lists of synonymous names and pseudonyms, and finally convert the names in a harmonized presentation format ("last, first"). After harmonizing the author names we generated visualizations of author life years, which are also listed in the library catalogue. In some cases this revealed ambiguous names that in fact referred to different authors with the same name but who lived at different times (Figure 2). We therefore used the combined author name and life year as the final unique identifier for each author, and removed names that could not be unambiguously identified from the final data so as to avoid bias. Ultimately, we generated lists of the most commonly accepted author names, and also of the discarded names to monitor the conversion quality and to spot any obvious errors in the data handling: these summary tables are publicly available at https://github.com/rOpenGov/estc/blob/master/inst/examples/summary.md. Every detail of this analysis is fully transparent, and any observed errors can be fixed in the source data and algorithms, and this iterative process continues until the majority of the names are handled correctly by the analysis ecosystem. Whereas the original data lists the author names for 22,320 documents, we were able to find a unique, unambiguous author name for 83 per cent (18,493) of these documents after the pre-processing. Similar conversions take place for the publication places and years, document dimensions and other fields: the full algorithmic details can

be browsed at https://github.com/ropengov/estc. After this initial polishing, the database is ready to be subjected to final statistical analysis.

A look at the most common authors who wrote history based on the title count reveals a mixture of pamphleteers and writers who are more commonly understood as historians (Figure 1). We highlight three of these authors in Figure 1 for further comparison (William Prynne, Daniel Defoe and David Hume). We also took into consideration a namesake of the more famous eighteenth-century David Hume to underline the relevance of individuating the authors in the catalogue (Figure 2). It is noticeable that the birth dates of the most popular authors are fairly evenly distributed throughout the early modern period, indicating that there were no particular peak moments for publishing history titles by known authors.

It is useful when evaluating the nature of a particular author's works to set the number of titles on a timeline (Figure 3). What is noticeable in the comparison

Fig. 1: Early modern authors[17] who published the most titles on history according to the ESTC catalogue data; the highlighted authors are compared further below.

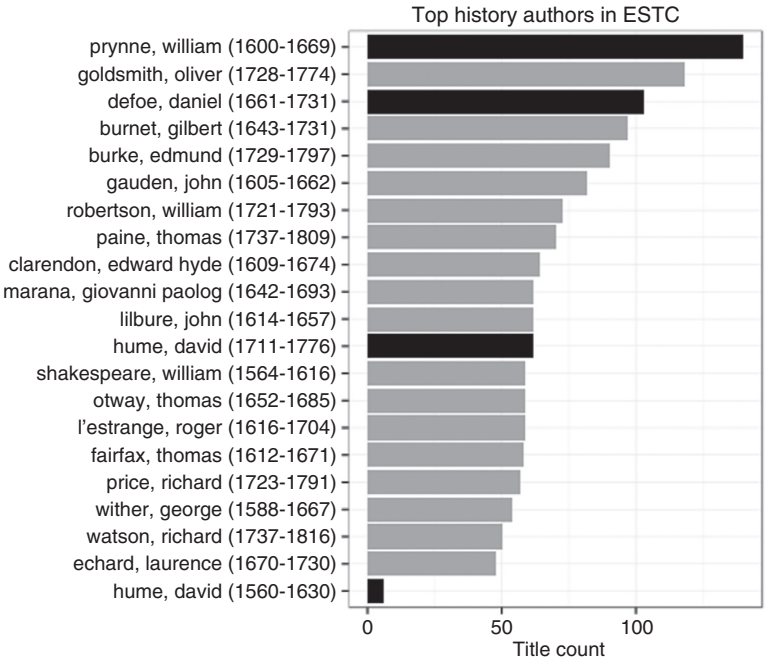
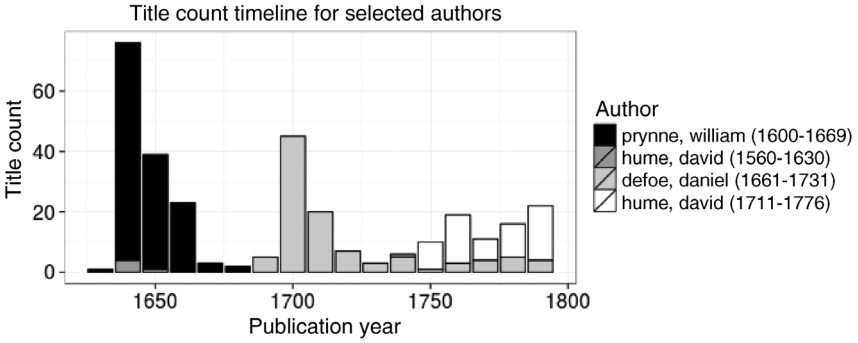


Fig. 2: The life spans of the top early modern authors based on the title count: the visualization also reveals ambiguities arising from authors having the same name but living at different times (e.g. David Hume).

Fig. 3: The title counts per year for William Prynne, Daniel Defoe and David Hume (highlighted in Figures 1 and 2) provide an overview of their publishing activity up until 1800.



of Prynne, Defoe and Hume is that William Prynne caused quite intensive peaks in publication numbers during the English civil war, but his works were no longer published after the late seventeenth century. Defoe caused a very steep peak in publication numbers in the Union debates (1705–1706),

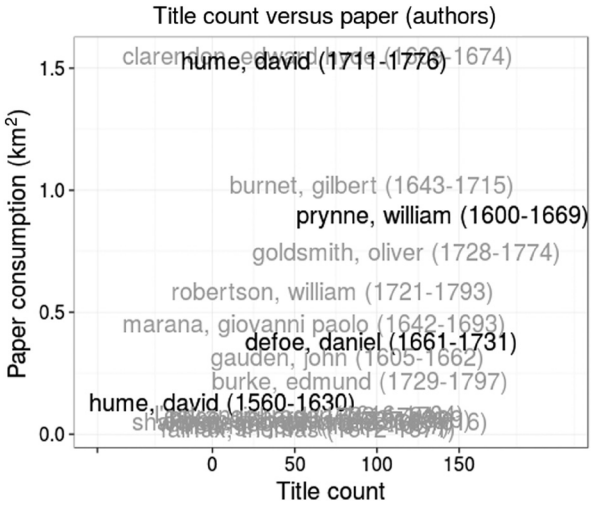
but his works continued to be published throughout the eighteenth century. Hume's historical writings showed a more steady development (there is no pamphleteering among them, as there is in Prynne and Defoe). Hume could be considered very successful in producing a steady flow of historical works throughout the second half of the eighteenth century.

Analysis of the document dimensions reveals further information on specific authors: this is another example in which considerable polishing of the original data fields is needed before proper statistical analysis is possible. We set up automated algorithms to recognize and harmonize the most commonly used forms of the standard document sizes, such as quarto, which is also commonly referred to as 4 to and 4º. In some cases the physical dimensions (in cm) are declared instead of the standard sizes. Where possible, our algorithms aim to augment such missing information based on ready-made conversion tables that assign the common standard sizes with their corresponding physical dimensions (see e.g. https://github.com/rOpenGov/bibliographica/blob/master/inst/extdata/documentdimensions.csv). Finally, standardized document-size estimates are obtained for most documents, facilitating the comparison of publication activity among different authors.

Figure 4 compares the top authors based on the title count in relation to the paper consumed in their books. From this perspective David Hume appears to have been successful indeed as a historian in terms of producing a steady flow of books of significant size. At the same time, Defoe's historical documents seem to be of more of a pamphleteering nature than those of William Prynne. Clarendon, Robertson, Goldsmith and Burnet also start to stand out in this graph, as anyone familiar with the historiography of early modern Britain might expect. Thus, when the evidence from the three previous graphs is combined a certain consistency in David Hume's historical publications emerges. Once he started publishing on history in the 1750s the volume grew steadily, and there was constant reproduction throughout the rest of the eighteenth century. Unlike Hume's writing, Daniel Defoe's works in particular are more random: although a prolific prose writer, he was also a pamphleteer switching from one topic to another. What is noticeable in William Prynne's publications is that he was very resourceful as a seventeenth-century writer on history, yet, after his death his works stopped being reproduced.

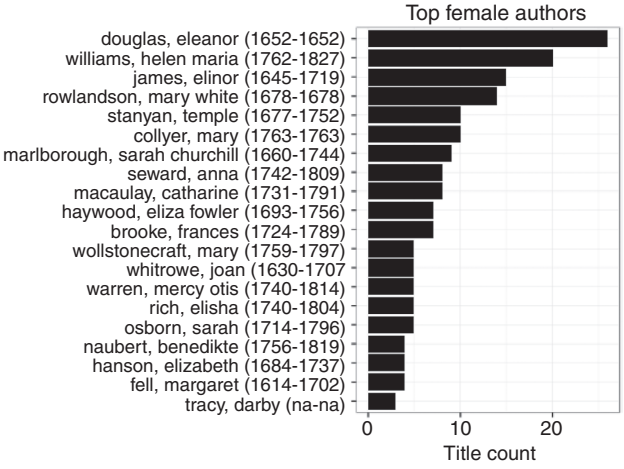
Analysis of the author-gender distribution gives an example of supplementing the original library catalogue data. We supplemented the author

Fig. 4: Title count versus paper consumption among the highlighted authors: the visualization reveals the nature of the authors' publications, distinguishing pamphleteering (many titles, few pages) and the authoring of books (fewer titles, more pages).



information by estimating the gender of each one on the basis of publicly available information on first names and genders from the US national census, the R package *gender* and other sources that is incorporated into our analytical ecosystem (https://github.com/rOpenGov/bibliographica/tree/master/inst/extdata/names). Although a significant proportion of female authors in the early modern period wrote under a masculine pen name, or anonymously, there were still a significant number of female authors writing history catalogued in the ESTC (Figure 5).[18] Even when women do not feature among the most prolific early modern authors of history, there were female authors with a significant volume of publications who compare favourably with the most famous male authors. In the future we hope to incorporate further data on pseudonym genders and on variations in name-gender distributions over time. Such improvements are easily incorporated into our ecosystem, and our analysis provides the first quantitative estimate of the publishing activities of female authors throughout the study period; we anticipate that the overall trends will remain largely robust for such updates.

Fig. 5: The most active known female authors based on the title count: the gender is inferred automatically from the first names.[19]

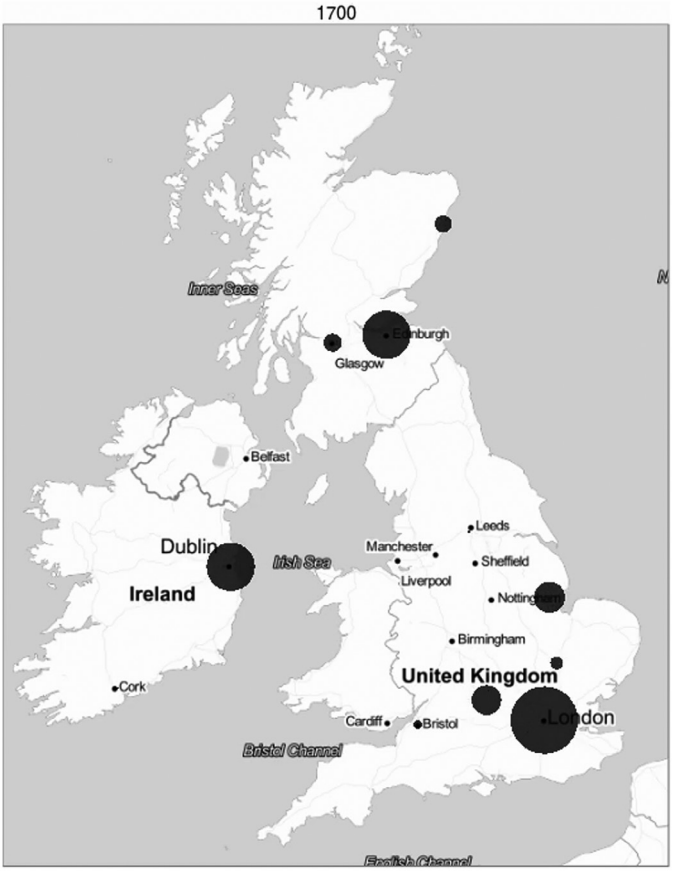


5. Where was history published?

London dominated the publishing business in Britain and North America during the early modern period until 1800 (Figure 6: see Myers, 1973; Pollard, 1978; Twyman, 1994).[20] Most importantly, the Licensing Act restricted publishing in the English provinces until 1695, although booksellers played a crucial role in distributing books in peripheral areas while printing was limited (Barnard & Bell, 2002; Feather, 2004). It is also well known that the top publication locations shown in Figure 7 (Dublin, Edinburgh, Philadelphia, Boston and the University towns of Oxford and Cambridge) were also of importance once the book trade became more accessible.[21] What has not been possible without great effort thus far is the quantitative comparison of different, especially smaller, publication locations, which would facilitate investigation into how this might reflect the publication of different genres. Below we analyse the major historical trends in publishing outside of London, in particular in Ireland, Scotland and the USA (the three biggest producers of printed documents after England).

A comparison of paper consumption and title count reveals, for instance, that there were relatively more publications compared to overall paper

Fig. 6: Publication volumes at the six top publication locations in Britain and Ireland, year 1700: the circle diameter corresponds to the logarithm (log10) of the title count.[22]



consumption in the area what we now know as USA than in the other two countries (Figure 8). Further comparison among the three during the early modern period reveals that publishing activity in the US was very sizeable in terms of the number of titles, but in terms of paper consumption the volume seems to have been proportionally much lower than in Scotland and, especially, Ireland (Figure 9). The implication is that many of the historical documents published in the USA were pamphlets.[23] Given the volume development in US publication especially towards the end of the eighteenth

Fig. 7: The top publication locations in Britain and North America ranked by the title count (number of published titles).

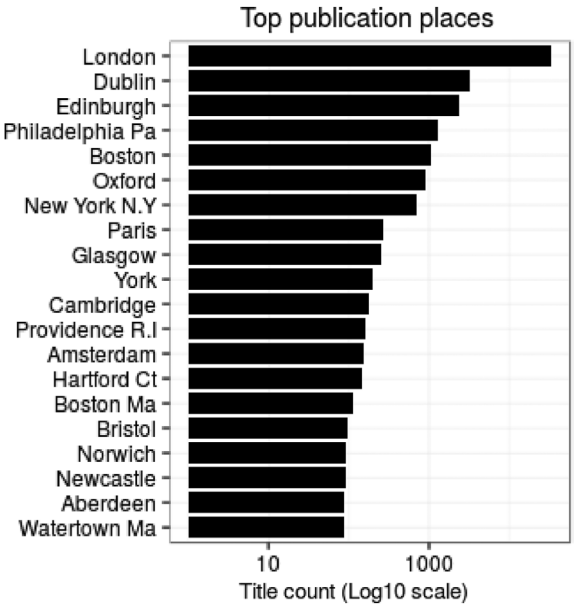


Fig. 8: Title count and overall paper consumption in the top publication locations: Places in US in darker colour.

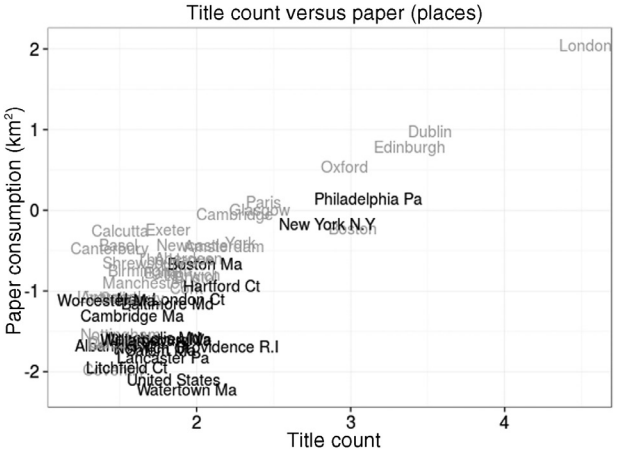
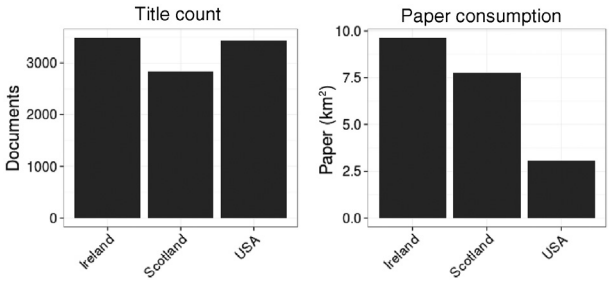


Fig. 9: Title count and paper consumption in Ireland, Scotland and the USA.[24]

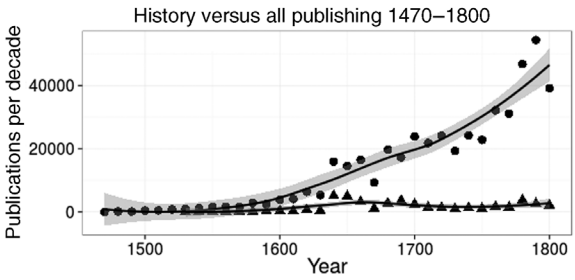


century, this gives an intriguing picture of the historical relevance of pamphleteering—in some sense the young colonies were going through much of what Britain had experienced in the seventeenth century. Meanwhile the tendency, especially in Scotland and Ireland, was to publish octavo-sized books.

6. How did publishing on the subject of history change over time?

Despite the advancements in historiography, there is still uncertainty about disciplinary boundaries and the adoption of a quantitative approach to the subject.[25] We investigate the seemingly naïve question of what history is in relation to publication volumes, taking as our starting point all documents that contain the word 'history' in their classification field in the ESTC catalogue. It is evident from a comparison of publication volumes in history and all documents in the ESTC that other subject areas expanded much more rapidly (Figure 10). This might seem counterintuitive to the assumed relevance of the rise in historical awareness and the concept of progress in the late eighteenth century: one might rather assume that, proportionally, there would have been an increase in publishing on history towards the end of the eighteenth century. A more detailed look at the title count of documents including the word 'history' reveals substantial peaks in the numbers of titles published in the 1640s, the 1650s and the 1690s, attributable in part to the existence of the Thomason tracts, and also to a rise in the numbers of pamphlets published during those times.

Fig. 10: A comparison between the title count for history publications and for all documents in the ESTC catalogue, 1470–1800.



Analytical bibliography is not, of course, about counting pages (Tanselle, 2000).[26] However, it may prove useful for estimating paper consumption. We have enough information, for instance, to distinguish three-volume works in folio from a half-sheet broadside based on our analytical ecosystem.[27] We can study book production by looking at book sizes and, with regard to the ESTC even by providing exact quantitative estimates on paper consumption for each year of this 300-year period during which the publishing system was established. The analysis of paper consumption requires information on document dimensions, page counts, and print-run sizes. Although none of these is available in the ESTC catalogue in a directly usable format, we can derive estimates based on the available information that can be extracted from the data fields via dedicated functions that we have implemented for this purpose. The cleaning up of the document dimensions is described in Section 4 above. Estimating page counts requires the summing up of information on cover pages and special pages, actual content, and possible multi-volume information, for example, according to the standard rules for page listing. The functions in the estc and bibliographica R packages can interpret the page-count field and convert it into exact numeric estimates of the total page count in each document. We have also added specific unit tests to check automatically that standard examples are converted correctly whenever the algorithms are updated. For print-run sizes we use the estimate of a rough 'London average' of 1,000 copies for every edition regardless of the format.[28] It is well known that there is variation in actual print runs, and that the numbers rise considerably in times of crisis, especially regarding the most popular pamphlets (even up to 10,000). As a general rule it seems to apply fairly consistently that the print runs ranged from 750 to 1,250 (or 1,500 copies maximum). When an edition was sold out a new

one followed, and we can make quite good estimates of the number of copies sold, especially of books, by counting the number of editions. In cases of missing page-count information we have used averages calculated over books of a similar size, treating multi-volume sets as a separate category. We have also manually checked that the amount of missing information does not change significantly between historical periods and thus bias the analysis. There will also be cases of lost books, but because our approach does not rely on having a complete corpus, this will not form a bottleneck in terms of reaching general statistical conclusions.[29] We could incorporate further perspectives to improve the estimates of overall paper consumption, such as taking supplementary information from the *Early English Booktrade Database* and other sources into our data-analytical ecosystem. In short, we have used the ESTC catalogue as a basis for an ecosystem where the idea is that this can then be supplemented from other information sources. In an excellent study Gants (2002) provides what he calls a snapshot of five years in the London book trade, examining in accurate detail the question of how much paper was used in sheets in London publishing. Whereas Gants measures the amount of paper used in making the books, we look at the volume of paper in the books recorded in the overall ESTC catalogue. We are not concerned with what might have been trimmed off these books, for example.

A comparison of the total number of history titles (Figure 11) with our estimates of overall paper consumption in the same documents (Figure 12) shows that the overall volume of history publishing in fact rose rather sharply towards the end of the eighteenth century. The indication is that many more books were published than pamphlets, which had previously dominated publications on history.[30] This finding supports the notion that the relevance of historical analysis did indeed take a new turn towards the end of the eighteenth century. Thus, whereas the number of history titles published annually over time remained stable, the publication volume measured in paper consumption rose exponentially during the eighteenth century.[31] One might assume that the exponential growth in paper consumption was even more substantial if all the documents in the ESTC are taken into consideration.[32] Simon Eliot (2007) states, 'the explosion of book production and of all kinds of print production actually took place in the nineteenth century, and more precisely after 1850, after what has long been called the "industrial revolution."' This might be true, but at the same time, in terms of the actual volume of paper usage, the eighteenth-century development could be regarded as a genuine outburst during the handpress period.[33] This also makes sense with

Fig. 11: The title count of history publications, 1470–1800.

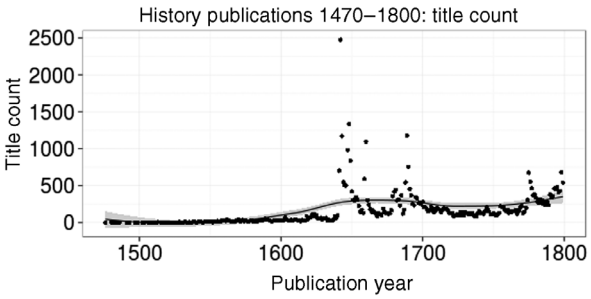


Fig. 12: Paper consumption in history publications, 1470–1800.

regard to the technological innovation of machine presses in printing that eventually ended the handpress period in the 1830s. Our analysis suggests that handpress printing was pushed to its limits during the latter part of the eighteenth century, and that technological innovation was therefore called for.

It is commonly assumed that the average size of book formats became smaller towards the end of the seventeenth century for practical reasons.[34] This complies with our finding that more books on history were published during the eighteenth century. In this regard, too, 15th- and 16th-century publications tend to be more sizeable than 18th-century documents. A good example of this is Holinshed's *Chronicles,* which shows as a clear publication peak in the 1570s according to the data (Figure 13). Later on, even though the folio size was the most common for books until the end of the seventeenth century in the sample of history documents studied for this article, history publishing involved fewer heroic undertakings than the *Chronicles*.

Fig. 13: Average paper consumption per document in history publications, 1470–1800.

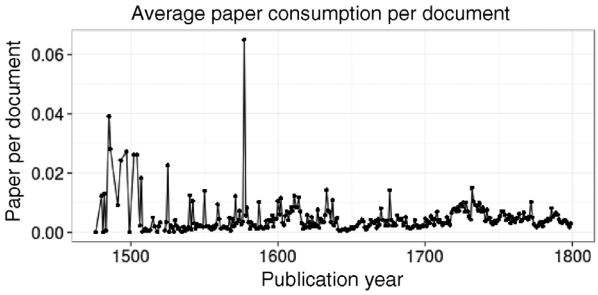
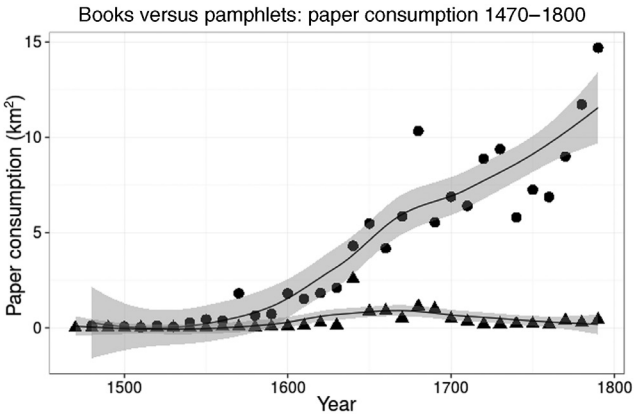


Fig. 14: Paper consumption in books versus pamphlets, 1470–1800.



In terms of book history, this article moves beyond the mere counting of titles and pages. It is obvious that both pamphlets and books played an important role in early modern publishing (Halasz, 1997; Raymond, 2003), and it is relevant to make a distinction between the two in studying publishing trends. When they are compared on the basis of the page count, as we have done (Figure 14), it is evident that although the average size of an individual book was larger during the earlier stages in the history of publishing, annual paper consumption rose steadily in the late seventeenth century.[35] The octavo form was commonly used for books on history in the later part of the eighteenth century (Figure 15). We can also point to the time when folio-sized publications begin to decline. Also the increased number of pamphlets caused by the collection of variants in the Thomason Tracts included in the ESTC is clearly

visible in the paper consumption of quarto-sized documents during the Civil War era.[36] It is thus clear that the octavo book hailed a new form of publishing in the eighteenth century. Intriguingly enough, a title-count-based comparison of the octavo and folio publications of top authors (Figure 16) reveals

Fig. 15: Paper consumption for different book formats over time.

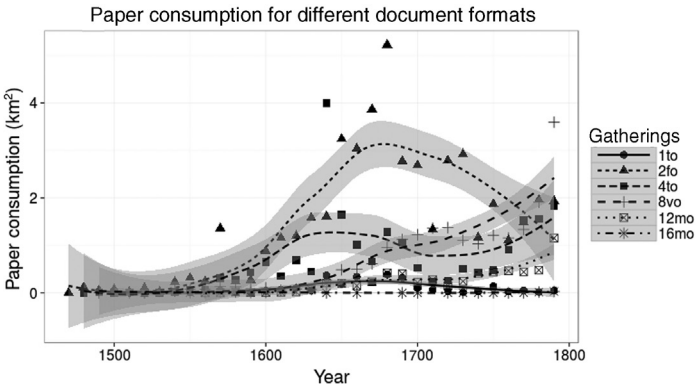
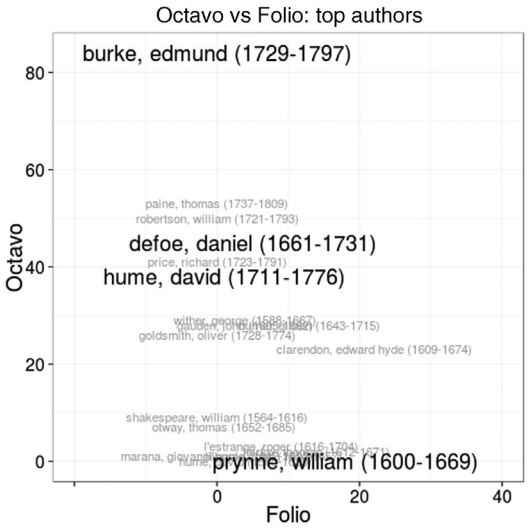


Fig. 16: A comparison by title count between the octavo and the folio format among the top authors (see Figures 1 and 2).



that Edmund Burke published the most octavo volumes. William Prynne, a seventeenth-century author, had no octavo publications at all, unlike some of his contemporaries (such as Shakespeare) who were also published in the eighteenth century. David Hume, whose overall output consumed the most paper based on our analysis, turns out to have a moderate number of publications in the octavo format.

Specific peculiarities, such as the presence of the Thomason Tracts in the mid-17th century, have to be taken into account when interpreting the library catalogues. At the same time we can infer that also real social change is reflected in the publication volumes. We consider this a more direct way of accounting for what has been considered crucial in book history since Elizabeth Eisenstein's (1980) contributions to the field. For example, the English Civil War, as well as the Restoration, the Glorious Revolution and the Union debates of 1705–1706, are clearly reflected as increased numbers of published titles in the graph depicting publishing activity in Edinburgh (Figure 17). At the same time, American Independence did not seem to cause a publication peak in history titles in Edinburgh. This indicates that in the early modern period history was published locally during times that also turned out to have particular historical relevance in that particular place, suggesting a degree of subjectivity in the catalogue classifications.

Descriptive bibliography facilitates the analysis of books as material objects with particular shapes and sizes whose distribution and production had

Fig. 17: The publishing of historical works in Edinburgh on a timeline highlighting the eras of the English Civil War (1642–1651), the Restoration (1660), the Glorious Revolution (1688–1689), the Union Debates (1705–1706) and American Independence (1776).

strong direct and indirect effects on the development of the early modern commercial society. [For the vast literature on this subject, see e.g., Barber (1994); Baron, Lindquist, & Shevlin (2007); Bermingham & Brewer (1995); Febvre & Martin (1990), p. 155–166; Rich & Wilson (1977); Sher (2007); Steinberg (2001), p. 106–129.] Librarians may make mistakes in inserting data into the catalogues, and not everything is catalogued, hence library and information studies are open to correction and improvement (De Morgan, 1853). The research is lacking in such large-scale analysis, however, as the catalogues have been used merely to locate individual books. It is clear that the value of the metadata in library catalogues and the potential new perspectives it opens up have been underestimated.

7. Conclusion

We have demonstrated how library catalogue data can be used to analyse large-scale developments of knowledge production such as the publishing of historical writings in early modern Britain and North America. This is only the beginning. Library data facilitates analysis of the development of various documents as material objects. In the case of history, for example, the long-term transformation from the folio to the octavo format is indicative of the growing readership as well as the wider distribution of historical knowledge during the later eighteenth century. Moreover, the analysis of paper consumption as well as the number of titles published highlights different aspects of publication volumes and allows differentiation between books and pamphlets, which in turn facilitates the analysis of individual authors and their oeuvre and places of publication. The case of the US and the volume of pamphleteering in the field of history are indicative of the resonance of knowledge production with regard to social change, which is also evident in the publication activity in Edinburgh analysed above. Our automated and open source statistical tools constitute a promising starting point for similar studies developing quantitative, data-driven analyses covering the whole of early modern Europe, when work will begin on catalogues that really cross national boundaries. Although we believe that our analysis gives a robust picture of publication activity in the early modern era, the open-source approach guarantees that such a picture can only improve over time given that the data and algorithms are easily updated. Any detected errors or shortcomings can be reported via issue tracker and corrected, and the full analysis can be

updated automatically, taking the corrections into account in all the steps. It is evident that the book was the most important vehicle of knowledge transfer in early modern Europe (Johns, 1998). Eventually, we believe, this kind of undertaking will be necessary to produce a more coherent account of historical constructions such as the Enlightenment and the European Republic of Letters. We anticipate that as libraries and other data holders give researchers and the general public full access to their data resources, there will be a rapidly increasing demand for open source ecosystems of data analysis in this field, the application of which we have demonstrated in this article.

References

Adams, T.R. (1980). *The American controversy: A bibliographical study of the British pamphlets about the American disputes, 1764–1783*. Providence: Brown University Press.

Alston, R. (1984). The British Book Trade, 1701 to 1800. *Publishing History*, *16*, 43–86.

Alston, R. (2004). The history of ESTC. *Age of Johnson*, *14*, 269–329.

Altick, R.D., & Fenstermaker, J.J. (1993). *The art of literary research* (4th ed.). New York: Norton.

Amory, H., & Hall, D.D. (Eds.). (2000). *The history of the book in America. I. The colonial book in the Atlantic world*. Chapel Hill, NC: The University of North Carolina Press.

Aspinall, A. (1948). Statistical accounts of the London newspapers in the eighteenth century. *English Historical Review*, *63*, 201–232. doi: 10.1093/ehr/LXIII.CCXLVII.201.

Barber, G. (1994). *Studies in the booktrade of the European enlightenment*. London: Pindar Press.

Barnard, J., & Bell, M. (2002). The English provinces. In J. Barnard & D.F. McKenzie (Eds.), *The Cambridge history of the book in Britain*, vol. 4 (pp. 665–686). Cambridge University Press.

Baron, S.A., Lindquist, E.N., & Shevlin, E.F. (Eds.). (2007). *Agent of change: Print culture studies after Eliabeth L. Eisenstein.* Amherst, MA: University of Massachusetts Press.

Belanger, T. (1982). Publishers and writers in eighteenth-century England. In I. Rivers (Ed.), *Books and their readers in eighteenth-century England: New essays* (pp. 5–25). Leicester University Press.

Bell, M. (2001). A quantitative survey of British book production 1475–1700. In L. Hellinga (Ed.), *The scholar and the database: Papers presented on 4 November 1999*

at the CERL conference hosted by the Royal Library, Brussels (pp. 15–21). London: CERL. Retrieved September 10, 2015, from http://documents.cerl.org/publications/cerl_papers_ii.pdf.

Bell, M., & Barnard, J. (1992). Provisional count of STC titles, 1475–1640. *Publishing History*, *31*, 47–64.

Bell, M., & Barnard, J. (1998). Provisional count of Wing titles, 1641–1700. *Publishing History*, *44*, 89–97.

Bermingham, A., & Brewer, J. (Eds.). (1995). *The consumption of culture, 1600–1800: Image, object, text*. London: Routledge.

Coleman, D.C. (1958). *The British paper industry 1495–1860: A study in industrial growth*. Oxford: Clarendon Press.

Crawford, P. (1984). Women's published writings 1600–1700. In M. Prior (Ed.), *Women in English society 1500–1800* (pp. 211–282). London: Methuen.

De Morgan, A. (1853). On the difficulty of correct description of books. In *Companion to the Almanac* (pp. 5–19).

Eisenstein, E. (1980). *The printing press as an agent of change*. Cambridge University Press.

Eliot, S. (1994). *Some patterns and trends in British publishing, 1800–1919*. London: Bibliographic Society.

Eliot, S. (1997). Patterns and trends and the *NSTC:* Some initial observations, part 1. *Publishing History, 42,* 79–104.

Eliot, S. (1998). Patterns and trends and the *NSTC:* Some initial observations, part 2. *Publishing History*, *43*, 71–112.

Eliot, S. (2002). Very necessary but not quite sufficient. A personal view of quantitative analysis in book history. *Book History*, *5*, 283–293. doi: 10.1353/bh.2002.0006.

Eliot, S. (2007). From few and expensive to many and cheap: The British book market 1800–1890. In S. Eliot & J. Rose (Eds.), *Companion to the history of the book* (pp. 291–302). Hoboken, NJ: Wiley-Blackwell.

Feather, J. (1986). British publishing in the eighteenth century: A preliminary subject analysis. *Library*, *s6-VIII*, 32–46. doi: 10.1093/library/s6-VIII.1.32.

Feather, J. (2004). The history of the English provincial book trade: A research agenda. In B. McKay, M. Bell, & J. Hinks (Eds.), *Light on the book trade: Essays in honour of Peter Isaac* (pp. 1–12). London: Oak Knoll Press.

Febvre, L., & Martin, H.-J. (1990). *The coming of the book: The impact of printing 1450–1800* (transl. David Gerard). London: Verso.

Ferguson, M. (1996). Renaissance concepts of the "woman writer". In H. Wilcox (Ed.), *Women and literature in Britain 1500–1700* (pp. 143–168). Cambridge University Press. doi: 10.1017/CBO9780511470363.010.

Frank, J. (1961). *The beginnings of the English newspaper, 1620–1660*. Cambridge, MA: Harvard University Press.

Gants, D.L. (2002). A quantitative analysis of the London book trade, 1614–1618. *Studies in Bibliography*, *55,* 185–214. Retrieved September 10, 2015, from http://xtf.lib.virginia.edu/xtf/view?docId=StudiesInBiblio/uvaBook/tei/sibv055.xml;chunk.id=d42;toc.depth=1;toc.id=d42;brand=default.

Gaskell, P. (1972). *A new introduction to bibliography*. Oxford University Press.

Giles, P. (2001). *Transatlantic insurrections: British culture and the formation of American literature, 1730–1860*. Philadelphia: University of Pennsylvania Press.

Gillespie, R., & Hadfield, A. (Eds.). (2006). *The Oxford history of the Irish book, vol. III: The Irish Book in English, 1550–1800*. Oxford University Press.

Greetham, D.C. (1992). *Textual scholarship: An introduction.* New York & London: Psychology Press.

Greg, W.W. (1913). What is bibliography? *Transactions of the Bibliographical Society*, *12*(1), 39–53. doi: 10.1093/libraj/TBS-12.1.39.

Greg, W.W. (1956). *Some aspects and problems of London publishing between 1550 and 1650*, Oxford: Clarendon Press.

Greg, W.W. (1966). Entrance in the stationers' register: Some statistics. In J.C. Maxwel (Ed.), *W.W. Greg: Collected papers* (pp. 341–348). Oxford: Clarendon Press.

Halasz, A. (1997). *The marketplace of print*. Cambridge University Press.

Heawood, E. (1929). Sources of early English paper-supply. *Library, s4-X,* 282–307. doi: 10.1093/library/s4-X.3.282.

Heawood, E. (1930a). Sources of early English paper-supply. *Library, s4-X*, 427–454. doi: 10.1093/library/s4-X.4.427.

Heawood, E. (1930b). Paper used in England after 1600, I. *Library, s4-XI,* 263–299. doi: 10.1093/library/s4-XI.3.263.

Heawood, E. (1931). Paper used in England after 1600, II. *Library, s4-XI,* 466–498. doi: 10.1093/library/s4-XI.4.466.

Heawood, E. (1947). Further notes on paper used in England after 1600. *Library, s5-II,* 119–149. doi: 10.1093/library/s5-II.2-3.119.

Hills, R.L. (1988). *Papermaking in Britain, 1488–1988.* London: Athlone Press.

Hinks, J. (2012). The book trade in early modern Britain: Centres, peripheries and networks. In B. Rial Costas (Ed.), *Print cultures and peripheries in early modern Europe: A contribution to the history of printing and the book trade in small European and Spanish cities* (pp. 101–126). Leiden, Netherlands: Brill.

Hoftijzer, P.G. (2002). British books abroad: The continent. In J. Barnard & D.F. McKenzie (Eds.), *Cambridge History of the Book in Britain, vol. 4, 1557–1695* (pp. 735–743). Cambridge University Press.

Johns, A. (1998). *The nature of the book: Print and knowledge in the making*. University of Chicago Press.

Karian, S. (2011). The limitations and possibilities of the ESTC. *The age of Johnson*, *21*, 283–297.

Kraus, J.W. (1986). The history of publishing as a field of research for librarians and others. *Advances in library administration and organization, 5*, 33–65.

Lambert, S. (1981). The beginnings of printing for the House of Commons, 1640–1642. *Library*, *s6-3*, 43–61. doi: 10.1093/library/s6-3.1.43.

Lambert, S. (Ed.). (1984). *Printing for parliament, 1641–1700* (List and Index Society, Special Series 20). London: Swift Printers.

Lambert, S. (1987). The printers and the government, 1604–1640. In R. Myers & M. Harris (Eds.), *Aspects of printing from 1600* (pp. 1–29). Oxford Polytechnic Press.

Lewalski, B.K. (1993). *Writing women in Jacobean England*. Cambridge, MA: Harvard University Press.

Lindenbaum, P. (1995). Authors and publishers in the late seventeenth century: New evidence on their relations. *Library*, *s6-17*, 250–269. doi:10.1093/library/s6-17.3.250.

Mann, A. (1999). Embroidery to enterprise: The role of women in the book trade of early modern Scotland. In E. Ewan & M.M. Meikle (Eds.), *Women in Scotland c. 1100–1750* (pp. 136–151). East Linton: Tuckwell Press.

Martin, R.L. III (2007). North America and transatlantic book culture to 1800. In S. Eliot and J. Rose (Eds.), *A companion to the history of the book* (pp. 259–272). Hoboken, NJ: Wiley-Blackwell.

May, J. (1984). On the inclusiveness of descriptive bibliographies: Limitations of bibliographical catalogues like the *ESTC*. *Analytical and enumerative bibliography*, *8*, 227–238.

McKenzie, D. (1992). The economies of print, 1550–1750: Scales of production and conditions of constraint. In S. Cavaciocchi (Ed.), *Produzione e commercio della carta e del libro. Secc. XIII-XVIII* (pp. 389–425). Firenze: Le Monnier.

Myers, R. (1973) *The British book trade from Caxton to the present day: A bibliographical guide*. London: André Deutsch.

Myers, R., & Harris, M. (Eds.). (1997). *The Stationers' Company and the book trade 1550–1900*. Winchester and New Castle, DE: Oak Knoll Press.

Myers, R., & Harris, M. (Eds.). (1981). *Development of the English book trade, 1700–1899.* Oxford Polytechnic Press.

Myers, R., & Harris, M. (Eds.). (1982). *Sale and distribution of books from 1700.* Oxford Polytechnic Press.

Myers, R., & Harris, M. (Eds.). (1983). *Author/Publisher relations during the eighteenth and nineteenth centuries*. Oxford Polytechnic Press.

Phillips, J.W. (1998). *Printing and bookselling in Dublin 1670–1800: A bibliographical enquiry.* Dublin: Irish Academic Press.

Pocock, J. (1957). Review of *English Historical Scholarship in the Sixteenth and Seventeenth Centuries* by Levi Fox (ed.), Oxford University Press, 1956 and *A History of the Society of Antiquaries* by Joan Evans, Oxford University Press, 1956. *Cambridge Historical Journal*, *13*, 190–192.

Pollard, H.G. (1941–1942), Notes on the size of the sheet. *Library*, *4th series*, *22*, 105–137.

Pollard, G. (1978). The English market for printed books. *Publishing History*, *4*, 7–48.

Pollard, M. (1987). *Dublin's trade in books, 1550–1800*. Oxford University Press.

Raven, J. (2014). *Bookscape: Geographies of printing and publishing in London before 1800*, University of Chicago Press.

Raymond, J. (2003). *Pamphlets and pamphleteering in early modern Britain*. Cambridge University Press.

Rich, E.E., & Wilson, C. (Eds.). (1977). *The Cambridge economic history of Europe*, vol. V: *The economic organization of early modern Europe.* Cambridge University Press.

Roberts, J. (2002). The Latin Trade. In J. Barnard & D.F. McKenzie (Eds.), *The Cambridge History of the Book in Britain,* vol. 4, *1557–1695* (pp. 141–173). Cambridge University Press.

Sher, R.B. (2007). *Enlightenment and the book: Scottish authors and their publishers in eighteenth-century Britain, Ireland and America.* University of Chicago Press.

Snyder, H.L. & Smith, M.S. (Eds.). (2003). *The English short-title catalogue: Past, present, future.* New York: AMS Press.

Stallybrass, P. (2004). The library and material texts. *Proceedings of the Modern Language Association of America*, *119*, 1347–1352. doi: 10.1632/003081204X17914.

Stanton, J. (1988). Statistical profile of women writing in English from 1660 to 1800. In F.M. Keener & S.E. Lorsch (Eds.), *Eighteenth-century women and the arts* (pp. 247–254). New York: Greenwood Press.

Steinberg, S. (2001). *Five hundred years of printing* (ed. Rev. by John Trevitt). London: Oak Knoll Press.

Suarez, M.F. (2003–2004). Historiographical problems and possibilities in book history and national histories of the book. *Studies in Bibliography*, *56*, 141–170. Retrieved September 9, 2015, from http://xtf.lib.virginia.edu/xtf/view?docId=StudiesInBiblio/uvaBook/tei/sibv056.xml;chunk.id=d33;toc.depth=1;toc.id=d33;brand=default.

Suarez, M.F. (2009).Towards a bibliometric analysis of the surviving record, 1701–1800. In M.F. Suarez & M.L.Turner (Eds.), *The Cambridge history of the book in Britain*, vol. 5 (pp. 37–65). Cambridge University Press.

Suarez, M.F. (2015). Book history from descriptive bibliographies. In L. Howsam (Ed.), *The Cambridge companion to the history of the book* (pp. 199–219). Cambridge University Press.

Summit, J. (2000). *Lost property: The woman writer and English literary history, 1380–1589*. University of Chicago Press.

Tanselle, G.T. (1974). Bibliography and science. *Studies in Bibliography*, 27, 55–89. Retrieved September 10,2015, from http://xtf.lib.virginia.edu/xtf/view?docId=StudiesInBiblio/uvaBook/tei/sibv027.xml;chunk.id=vol027.02;toc.depth=1;toc.id=vol027.02;brand=default.

Tanselle, G.T. (1976–1977). Bibliographers and the library. *Library Trends*, 25, 745–762. Retrieved September 10, 2015, from https://www.ideals.illinois.edu/bitstream/handle/2142/6932/librarytrendsv25i4d_opt.pdf?sequence=1.

Tanselle, G.T. (1981). *The history of books as a field of study*. Chapel Hill, N.C.: Rare Book Collection.

Tanselle, G.T. (1988). Bibliographical history as a field of study. *Studies in Bibliography*, *41*, 33–63. Retrieved September 10, 2015, from http://xtf.lib.virginia.edu/xtf/view?docId=StudiesInBiblio/uvaBook/tei/sibv041.xml;chunk.id=vol041.02;toc.depth=1;toc.id=vol041.02;brand=default.

Tanselle, G.T. (2000). The concept of format. *Studies in Bibliography*, *53*, 67–116. Retrieved September 10, 2015, from http://xtf.lib.virginia.edu/xtf/view?docId=StudiesInBiblio/uvaBook/tei/sibv053.xml;chunk.id=vol053.02;toc.depth=1;toc.id=vol053.02;brand=default.

Thomson, A.G. (1974). *The paper industry in Scotland, 1590–1861*. Edinburgh: Scottish Academic Press.

Twyman, M. (1994). Two centuries of printing: Book production history diagrams. *Publishing History*, *36*, 103–114.

Veylit, A. (1994). *A statistical survey and evaluation of the "Eighteenth-century short-title catalog"*, unpublished thesis. University of California Riverside.

Weedon, A. (2007). The uses of quantification. In S. Eliot and J. Rose (Eds.), *A companion to the history of the book* (pp. 33–49). Hoboken, NJ: Wiley-Blackwell.

Wheeler, W.G. (1978). The spread of provincial printing in Ireland before 1850, *Irish Booklore*, 4, 7–19.

Notes

[1] This project has been financially supported by the Faculty of Arts at the University of Helsinki and National Library of Finland and Digitalia in Mikkeli. Leo Lahti has been partially supported by Academy of Finland (grant 256950). We would like to express our gratitude also to Kaius Sinnemäki, Hege Roivainen, Maija Paavolainen and Krister Lindén for their help in various different ways. An earlier version of this paper was presented at the Liber 2015 conference in London. We would like to thank our audience, and especially Marian Lefferts of CERL.

[2] For more about the ESTC (and the historical development of the catalogue and the difference between the English and the Eighteenth-Century Short Title Catalogues), see http://estc.ucr.edu/estcdean.html including the bibliography noted there.

[3] This article is planned as the first in a series of publications that will extend to the analysis of European knowledge production in general based on the ecosystem emanating from our work on different library catalogues.

[4] This initiative for this paper came from the agenda outlined by Michael Suarez (2003–2004), and especially what he states about the use of numbers and tools such as the ESTC (including the limitations) on p. 166.

[5] See, e.g., 'Finding the Text: Enumerative and Systematic Bibliography', in Greetham (1992), pp. 13–46 and 'Finding Materials' and 'Libraries', in Altick and Fenstermaker (1993), pp. 155–204.

[6] https://github.com/rOpenGov/estc. We would like to thank the British Library for providing us with the data used in this article.

[7] We are simultaneously working on many different library catalogues in an open data project. For more information on the research data and the latest advances, see: https://github.com/rOpenGov (estc, fennica, kungliga).

[8] For a good analysis and example of how to begin to use the ESTC catalogue in the manner that we develop here, see also http://douglasduhaime.com/blog/mapping-the-early-english-book-trade.

[9] On the development of the ESTC, see Snyder and Smith (2003) and Alston (2004).

[10] For a text-mining project on David Hume's *History of England*, see http://www.crassh.cam.ac.uk/events/25988.

[11] Veylit (1994) is perhaps the most comprehensive undertaking thus far. At the same time, it is quite telling that although there is a chapter on the digital book in *The Cambridge Companion to the History of the Book*, first published in 2015, there is no reference to the statistical (or quantitative) approach in the index.

[12] See especially Eliot (2002). For Eliot's own work on quantitative analysis, see Eliot (1994, 1997, 1998). For more optimism within the community of people taking a quantitative approach to book history, see e.g., Bell (2001).

[13] There are plans to update the ESTC in the form of ESTC21, with improved linked data features and richer data; see https://estc21.wordpress.com/collecting-data/.

[14] On the relevance of descriptive bibliographies, see Suarez (2015).

[15] https://github.com/rOpenGov/estc.

[16] On authorship in general, see Belanger (1982), Myers and Harris (1983) and Lindenbaum (1995).

[17] Of ancient authors, Flavius Josephus (37 – c. 100) and Caesar, Julius (100 BC – 44 BC) would have made the list of most published titles in ESTC, but they have been left out of this study for the sake of focusing on early modern authors.

[18] On the essential and growing amount of scholarship on female authors, see e.g., Summit (2000), Crawford (1984), Ferguson (1996), Lewalski (1993), Mann (1999) and Stanton (1988).

[19] Darby Tracy is a pseudonym, we have left her in the figure to highlight that there are many similar cases in the data.

[20] The role of the Stationers' Company is obviously of the highest relevance in any analysis of London publishing: see Myers and Harris (1997) and Greg (1956, 1966). On provincial book production, see especially Hinks (2012). It is also noticeable that, other than the Latin trade, the number of British books printed on the continent was very high throughout the early-modern era. On the relationship between British books and the continent, see also Hoftijzer (2002).

[21] On early modern US printing, see Martin (2007) and Amory and Hall (2000). On Dublin, see Pollard (1987) and Phillips (1998). On Irish provincial printing, see Wheeler (1978), and Gillespie and Hadfield (2006).

[22] You can also download the full video: https://github.com/rOpenGov/estc/blob/master/inst/examples/liber.mp4.

[23] See Adams (1980) for a detailed study of British pamphlets and late-eighteenth-century American disputes, and Giles (2001) for more on the eighteenth-century development of American literature).

[24] Some of the data for calculating the paper consumption in ESTC catalogue is missing so that this has an impact on this particular figure. We have estimated that approximately 24% of the information for US paper consumption is missing from this graph. The effect of this is such that the US bar might be up to 32% higher (the missing information does not effect Ireland or Scotland to any significant extent). Nevertheless, the US paper consumption is still considerably lower, even if all the data was available. The updated figures can be accessed on github when the data becomes available: https://github.com/rOpenGov/estc/blob/master/inst/examples/20151023-LIBER.md.

[25] In this sense, we are still battling with the same questions as posed in Pocock (1957).

[26] For an overview of bibliography, see Greg (1913), Tanselle (1974, 1976–1977, 1981, 1988) and Gaskell (1972).

[27] A problem pointed out by Maureen Bell in an unpublished paper delivered at the British Library in 2006.

[28] On the use of the London average print run, cf. Raymond (2003) p. 90 and the works cited there.

[29] One of the major shortcomings of the ESTC is the lack of data on eighteenth-century newspapers, which is why we have excluded most of the later newspapers from this study. On the statistical approach to eighteenth-century newspapers, see Aspinall (1948).

[30] For general studies on the development of the British book trade in the eighteenth century, see Myers and Harris (1981, 1982), Alston (1984) and Feather (1986).

[31] Unfortunately we cannot provide a paper-consumption comparison for all the documents in the ESTC because we have not yet obtained access to this data from the British Library.

[32] We have been waiting for some time to see if it is possible to access the full ESTC data from the BL to complete this analysis.

[33] On the British paper industry, see Coleman (1958), Hills (1988) and Thomson (1974).

[34] On paper and its uses in Britain, see Pollard (1941–1942) and Heawood (1929, 1930a,b, 1931, 1947).

[35] This graph clearly also reflects the economic aspects of printing: see e.g., McKenzie (1992).

[36] On printing in general during the civil-war era, see Lambert (1981, 1984, 1987) and Frank (1961). One aspect for further study, apart from the effect of duplicates caused by the presence of the Thomason Tracts, concerns the Greek and Latin works imported from the European continent. It is obvious that not all of the 'Latin trade' is included in catalogues such as the ESTC. This is not really a problem in our study given that our aim is to focus on book publishing through the ESTC catalogue rather than the absolute objective reality, and thereby to present a reliable view of general trends and patterns. On the 'Latin trade', see Roberts (2002).