Genre Classification Workflow For the ESTC

Iiro Tiihonen[1], Kira Hinderks[2]

Abstract:

This article introduces an open-box workflow for labelling 94 percent of the English Short Title Catalogue (ESTC) with a unified genre classification scheme, as well as an approach to evaluating the classifications. As the ESTC covers most of the surviving books published in the Early Modern Anglosphere, the categorisation offers new opportunities for large-scale quantitative research on Early Modern book trade. Our evaluation process directly engages with the ambiguity of any genre labelling or annotation schemes of Early Modern books and highlights problematic boundaries between the categories. We also provide summary statistics about the genre-wise composition of the ESTC, demonstrate how the new data can be used to detect biases in other data sets of Early Modern books and discuss further possibilities in genre-related computational work with the ESTC.

Keywords: book history, classification, computational history, digital humanities, evaluation process, workflow

Abstrakti:

Tämä artikkeli esittelee avoimen työvuon 94%. ESTC-tietueista kategorisointiin yhtenäisellä luokittelujärjestelmällä, sekä lähestymistavan näiden luokittelujen arviointiin. Kategorisointi mahdollistaa uusia laaja-alaisia analyyseja, sillä ESTC kattaa suurimman osan Britti-imperiumissa tai englanniksi varhaismodernina aikana julkaistuista kirjoista. Arviointiprosessimme käsittelee ongelmia, joita varhaismodernien teosten luokitteluun väistämättä liittyy. Lisäksi esittelemme tilastoja ESTC:n genrejakaumasta, osoitamme uuden aineston höydyn muiden varhaismodernien kirja-aineistojen vinoumien tutkimisessa ja keskustelemme siitä, millaiset laskennalliset lähestymistavat ESTC-teosten genrejen analysointiin voivat olla tulevaisuudessa mahdollisia.

Avainsanat: kirjahistoria, luokittelu, laskennallinen historia, digitaalinen historia, arviointiprosessi, työvuo

[1] University of Helsinki, ORCID: 0000-0003-0703-4556
[2] University of Helsinki, ORCID: 0009-0004-9233-9315

1. Introduction

The English Short Title Catalogue (ESTC) is a union catalogue[3] that covers most of the surviving books printed in the Early Modern Anglosphere. As books as material objects have survived exceptionally well[4] and the number of copies per edition varied far less than in the modern era[5], the data in the ESTC enables far more comprehensive and interpretable analysis of the objects that it represents than what is possible with most historical data sets. This article presents a workflow for classifying 94 percent of ESTC records with a unified categorisation scheme, describes how the resulting data was evaluated and summarises useful information about the genre distribution of the ESTC and its subsets. We devote significant attention to systematic close reading of the labels suggested by the workflow and demonstrate how errors and ambiguities can offer historically and methodologically valuable insights. We add new and comprehensive evidence for the argument that the proportion of religious publications declined over time, find a permanent jump in the proportion of political publications during the English Civil War that warrants closer scrutiny, and systematise the analysis of biases in Eighteenth Century Collections Online, a significant subset of the ESTC.

Much of the information in the ESTC (publication year, publisher information, physical description of the edition etc.) has already been extensively processed to a form suitable for data-analysis, most notably by the Helsinki Computational History Group (COMHIS) (Lahti et al. 2019). The workflow described here is a recent addition to this long-term effort of COMHIS. The HPC-HD project, a collaboration between COMHIS and several research groups of computer science, has significantly extended the large-scale analysis of language in subsets of the ESTC (most notably ECCO) with state-of-the-art machine learning methods, and produced computational approaches to various linguistic phenomena, genre included (Zhang et al. 2022). In addition to this wider context, an article that approached books as commodities that vary by type[6] (Tiihonen, Lahti, and Tolonen 2024) prompted the development of this workflow and made use of an earlier version of the outputs. This article introduces the (revised and improved) workflow for the first time, extends evaluation from a short discussion point in the appendix to a topic in its own right, and provides new information about the data that is the end result. As there is a plan

[3] A catalogue that covers the records of multiple libraries.

[4] The claimed survival rates (defined as at least only copy on an edition surviving and making it to the ESTC) vary from 55 percent for the period before the English Civil War (Hill 2018), to 70-75 percent for 1640-1700 (Raymond 2003; McKenzie 2002) and 90 percent for the eighteenth century (Suarez 2009). Despite the varying levels of evidence, assumptions and pure guesswork behind these numbers, it is at least safe to say that books as material objects have survived exceptionally well.

[5] There was significant variation in the number of copies printed per edition in the Early Modern English-speaking world (Jenner 2011; Watt 1990; Simmons 2002), but less so than in the modern era. Roughly 1000 copies per edition is often presented as a typical number (Raymond 2011; Gants 2002; Hill 2018).

[6] e.g. almanacks and multi-volume dictionaries would have differed significantly in terms of price, scope of potential buyers and utility.

to release the genre data as part of the COMHIS-version of the ESTC in the future, the article also serves us documentation for the genre variables.

Our aim can be understood as being wide but relatively thin. The conceptualisation of genre is limited to a single main label per edition (a debatable choice as a unit of text in itself), taken from a pool of very limited choices, and we do not consider many forms of information like visual cues (e.g. images) or complex embedding representations of texts. As we will discuss in the article, many other conceptualisations and corresponding computational approaches could and should be implemented as well. But this simplicity also enables effective interpretation of both the process and the results. Moreover, our evaluation of the outputs enables us to not only verify that the labels are largely correct, but also to identify potentially fruitful directions for further computational work on Early Modern genres. Instead of aiming to establish a single permanent and fixed conceptualisation of genre, the workflow presented here is a particular tool for particular (but not irrelevant) applications, which hopefully helps in the development of other approaches.

Our workflow has four major advantages compared to preceding efforts to categorise editions in the ESTC. First, it covers virtually the entirety of the ESTC. Second, it is eclectic and makes use of the multitude of possible sources of information available, from a large pool of manually annotated documents to established conventions of Early Modern book titles (e.g. ‘sermon preached at X’) instead of aiming for one solution (of typically inferior trustworthiness) that should cover the entire population to be classified. Third, it is open-box. All operations related to classification can be communicated to an audience without great technical expertise. Our eclectic and open box approach to classification offers an alternative to increasingly employed black box methods that, in addition to their merits in being able construct complex decision processes (e.g. about genre), have special challenges in generalising reliably outside of the training data (Kapoor and Narayanan 2023).

Fourth, our approach to evaluation is tailored to suit the ambiguous nature of many eighteenth-century publications. Instead of having a ‘gold standard’ test set as is typical in machine learning or statistics, we manually evaluated a large sample of records for the relative quality of the label. In addition to clearly right (1) or wrong (0) classifications, we added the category of plausible (0.5), which, as it turns out, covers most of the ‘errors’ in the classified data. Our results highlight that certain genre and topic distinctions are not bugs but features of the Early Modern publishing landscape: documents at the intersection between politics, religion and law, for example, are inherent parts of the print culture of the era. Instead of forcibly resolving/smoothing historical genre ambiguities, we aim to quantify their size and their patterns. As analyses of historical bibliographic data are not limited to the ESTC, we hope that our workflow also motivates other researchers working with similar material to consider whether their needs could be served by a simple and

transparent workflow for classification and evaluation if they utilised material and humanities expertise rather than (or in addition to) latest computational methods.

This paper is structured as follows: Section 2 highlights the limitations of the ESTC's native genre classification scheme and discusses previous scholarly efforts to mitigate these limitations, e.g. by creating new classification schemes. This section also examines why the use of unsupervised machine learning models can be problematic for genre classification and stresses the need for detailed process descriptions in articles that introduce new classification schemes. Section 3 briefly explains the terminology necessary for understanding the makeup and structure of the ESTC and introduces our genre labelling scheme. Section 4 describes our classification workflow, namely the steps for creating category labels. Section 5 discusses our evaluation of the accuracy of the generated labels. This section also highlights difficulties and pertinent category overlaps, showing that what at first glance appeared to be an error was often a genuine reflection of Early Modern print culture, where genres were far less distinct compared to the twenty-first century. The sixth and final section analyses the genre distribution in the entire ESTC and offers suggestions for how the genre data and classification workflow introduced in this paper can be utilised by future research.

2. The Complexity of Genre: Previous Classification Schemes of the ESTC and its Subsets

Genre is a complex and multifaceted phenomenon and there is no one correct classification scheme that captures all its nuances in different contexts. The most appropriate scheme is highly dependent on material and research questions. For historical data, in almost all cases, it is better to use a classification scheme which tries to reflect as closely as possible how historical actors understood genre. However, the genre and subject categories present in Early Modern union catalogues are often based on library cataloguing schemes that were developed in the nineteenth century, such as the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LCC). These schemes were created first and foremost to fit the needs of libraries, which differ from the objectives of a scholar wishing to analyse genre distribution patterns in the Early Modern period. The DDC and LCC reflect nineteenth-century understandings of genre and may impose anachronistic categories on earlier periods or fail to capture genres and subjects that were prevalent in one period but fell out of fashion later. Taking existing classification schemes as-is—especially without a systematic evaluation of their biases—can therefore lead to the reproduction of fundamental issues.

Early Modern scholars are usually well-aware of the limitations of the genre classification schemes native to the ESTC and other union catalogues but may continue to use them because genres are not their main object of study or because a more historically accurate option is not available. For instance, one of the earliest

quantitative analyses of genre distribution in Early Modern Britain was based on genre categories taken from the DDC (Feather 1986). The reason for this was that Feather's dataset, the *Eighteenth Century British Books* (ECBB) short-title union catalogue, used the DCC. A rival catalogue to the ESTC, the ECBB was born out of the *Project for Historical Biobibliography* (PHIBB) at the University of Newcastle and was completed in 1981. Due to its smaller scope and more limited coverage, the ECBB has been dubbed as 'the poor library's ESTC' (Barr 1981, 229) and reviewers—some involved in the then-ongoing development of the ESTC—were only too keen to point out the ECBB's flaws, including the anachronistic elements present in the DCC (Snyder 1983; Amory 1983). In his quantitative analysis of the ECBB, Feather sought to curb the DCC's worst excesses by only including high-level subject categories. Despite Feather's best efforts, some issues remain: the category of 'social sciences', the second most frequent in Feather's analysis, is especially problematic. Its individual components are never explained, and the concept, although coined in France in the 1770s ('sciences sociales'), did not become a distinct tradition until the nineteenth century (see e.g. Baker 1964).

This type of anachronism is not only present in catalogues containing nineteenth-century genre classification schemes but also affects catalogues whose schemes were developed in the twentieth century. One such example is Eighteenth Century Collections Online (ECCO), the most extensive subset of the ESTC for which records have been classified by topic. ECCO covers roughly half of the eighteenth-century records of the ESTC and crucially includes digitised and machine-readable versions of the full text of the editions that it covers. Gale, the company that provides access to ECCO, has labelled all records in ECCO with its own classification scheme. Unfortunately, the classification scheme does not suit eighteenth-century records well, with many of the categories not corresponding with historically motivated genre or topic differences of the period.[7] The categories of 'Social Sciences' and 'Religion and Philosophy' are especially problematic in an eighteenth-century context, as they amalgamate very heterogeneous text types.

Scholars have sought to address this problem in multiple ways, with some mitigation strategies more successful than others. Under the assumption that a fully data-driven process can eliminate human biases, unsupervised machine learning methods for classification purposes have become increasingly popular: recent examples include the use of topic modeling on the HathiTrust Digital Library (Almelhem, 2023), EEBO (Grazl, 2023) and Goldsmith's Kress Collection (Tiihonen, 2022). The term 'topic model' itself is misleading, as it can lead users not familiar with the mathematics of topic models to assume that there is something inherently topic-like in the distributions generated by a topic model. This is not the case, as traditional topic models only generate distributions over units of text, and to our knowledge other unsupervised methods cannot make great ontological claims either. The general

[7] ECCO is also an unbalanced representation of the topic distribution of the ESTC (Tolonen, Mäkelä, and Lahti 2022), and we extend the analysis of this issue in the paper.

sensitivity of traditional topic models to particulars of data pre-processing and the dangers of ‘eyeballing’ interpretations of ‘topics’ (Shadrova 2021; Gillings and Hardie 2023) are highly undesirable qualities when working with historical data. It is hard to see how production of categories with unsupervised methods in general could fully escape these problems.

To better represent Early Modern conceptions of genre in a way that is most suitable for their own research questions, many scholars have created annotated subsets of the ESTC that focus on specific periods or certain groups of authors (Corns 1986; Gants 2002; Fielding and Rogers 2017; Hill 2018). However, not all publications report the extent to which—or even whether—harmonisation and unification have been performed. Among studies that introduce new classification schemes, a recurring issue is that descriptions of the workflow are sometimes absent or underdeveloped. Rationales for the creation of certain labels and an evaluation of the genre data quality in relation to the existing scholarship are also frequently omitted. For instance, Suarez (2009) proposes an 11-category genre classification scheme for the eighteenth-century part of the ESTC but does not explain whether this scheme is a modification of the original ESTC subject classification or whether it was developed from scratch. Suarez notes that this genre annotation was performed on a relatively small subset of the ESTC (ca. 24000 records) but does not systematically assess the coverage of this subset or the precision of his proposed genre categories. Such omissions make it more difficult for readers to understand which research questions can be answered with the help of the newly developed genre classification scheme.

There are two long standing scholarly traditions of developing and refining genre classification schemes in sync with other bibliographic work on the most important subsets of the ESTC: Early English Books (EEBO) at Lancaster University (e.g. Murphy 2019) and ECCO at the University of Helsinki (e.g. Zhang et al. 2022). Both traditions are examples of successfully integrating humanities domain knowledge with sophisticated computational methods. In contrast to unsupervised approaches, here, the humanities scholar is the prime mover of the process: for example, by manually annotating large data samples with genre labels, drawing on one’s domain knowledge and expertise. Providing a way of classifying genre in the whole ESTC, the workflow introduced in this paper builds on and contributes to this rich tradition.

3. Definitions and Taxonomies

Data Vocabulary

The article makes heavy use of a terminology designed to distinguish different types of data in the ESTC. For clarity, this section briefly introduces that terminology. The basic unit of the ESTC is a record representing one ‘printing operation’. That is, a set of copies printed of a given text as a part of the same process. In some more

complicated situations this definition and record do not perfectly correspond to each other.[8] In our vocabulary, records are called editions. The original ESTC records follow the MARC[9] format for bibliographic data.[10] This data has been harmonised, standardised, and enriched (Lahti et al. 2019) to a form more suitable for data-analysis. Of the raw MARC data fields, the classifications described here rely primarily on the fields that contain the title of an edition (245a and 245b), information about bibliographies that cite a given edition (510a), references to microfilm collections that include a copy of a record (533f), and those fields that contain keywords describing the genre, topic and format of an edition (600, 610, 650, 651, 655). The keywords and full titles were converted to lowercase text in the classification workflow.

In addition to editions, we sometimes refer to works. Work groups all editions with more or less the same textual content together. So, for example, there are several editions of the Bible in the ESTC with their distinct edition-level id, but they should map to the same work and corresponding work-id. The ESTC does not have comprehensive information about works as such, but this information has been algorithmically derived by the COMHIS group based on the title and author information in the ESTC (Ijaz, Roivainen, and Lahti 2019).

Among other information, ESTC editions have a title. Early Modern titles were often very long, even taking the first whole page of the publication. For this reason, they often also have a 'short title' (for which 'ST' in ESTC stands for) for easier communication. In the classification workflow, we used the longer version of the title, as it provided more information about the topic of an edition.

ESTC does not have a unified topic or genre classification, but many of the editions have information that is very useful for classifying them with one. Perhaps most importantly, there are a great number of keywords relating to genre, topic and format that are part of the information characterising records. When referring to topic, genre and format keywords in the ESTC, we mean this information. Additionally, editions in the ESTC often refer to bibliographies or microfilm collections that cite or contain photocopies of the edition. For example, the *Goldsmiths' Kress* microfilm collection (Whitten 1978) covers publications relevant for economic history and *History of English Drama 1660–1900: Volume 6, A Short-title Alphabetical Catalogue of Plays* by Allardyce Nicoll (1959) lists plays. When we refer to microfilm and bibliography collection information in the ESTC, we mean this kind of data.

[8] For example, multiple issues of a periodical are often stacked under the same record and multi-volume works were sometimes printed in pieces over a long period of time, and are sometimes still grouped together under a single record.
[9] https://www.loc.gov/marc/bibliographic/.
[10] The British Library hosts an interface to query the 'original' edition data, but at the moment it is down because of a cyberattack. A somewhat modified version of the data is available for querying at: https://estc.printprobability.org/

Labelling scheme

There are three hierarchical levels in our classification scheme. The supercategory, the main category, and the subcategory. Of these, the middle one **(main category)** is the most important one. It is a modest modification of the categorisation scheme used in previous articles of the Helsinki Computational History Group (COMHIS) (Zhang et al. 2022; Ryan and Tolonen 2024), that was also used to classify Early Modern editions of the ESTC. The categorisation scheme is the result of a thorough investigation of all of the known books published by Andrew Millar, a prominent eighteenth-century publisher. Rooted in scholarship about Early Modern books and designed for the era and location (eighteenth-century British Empire) that make most of the ESTC data, we adopted it and mainly left it untouched.

The most important difference to this scheme is that, in our scheme, the physical size of the edition affects its classification (ephemera-entertainment and ephemera-practical only include short editions), and that the categories of literature and art have been merged (literature and arts). In this sense it goes even further than most genre-like classification schemes (in comparison to strictly topic-based classifications), as it also considers the material cues (physical size) given by the publication about its category.

The reasoning behind the size-related classifications is that in some cases the size is a part of the publication's type. For example, short publications that dealt with scientific matters like astronomy tended to lean more towards immediate practical uses (e.g. almanacks) than longer works with superficially similar contents. As for literature and arts, we ended up concluding that it was too difficult and often not meaningful to distinguish between texts related to performing arts (poetry, plays) and prose, and they are often convoluted in eighteenth-century publications (e.g. different kind of miscellanies that contain both), making the distinction very artificial. The subcategories aim to compensate for this merge by specifying the type of publication (e.g. play, poem or novel) with more precision when possible.

Main category	**Includes the following publication types**
religion	Bibles, sermons, prayers, theology, religious debates (can also be politics in some instances), psalms and hymns.
literature and arts	Periodicals, novels, stories, visual arts, plays, poems, songs, pictures, operas, satires.
ephemera-entertainment	Chapbooks, broadside poems. Short items (pamphlets) that would otherwise be included in literature and arts or history.
politics	Political pamphlets, societal discussion, news about wars, foreign politics etc (note: periodicals are still in literature), commerce, social issues (e.g. poverty).
law	Legal and administrative documents, legal theory. As a rule of thumb, commentary and debate of legislation to politics; and description and statement to law.
science	Mathematics, natural sciences, agriculture, technology, military organisation and building,

	logistics (e.g. canals etc.), natural history, travel literature. Overlaps with literature: description of facts is more towards science and satirical/and or entertaining purposes towards literature.
education	Manners, upbringing manuals to do X, grammars, primers, dictionaries, conduct of life. Overlaps with topic-related categories, (e.g. introduction to mathematics).
philosophy	Logic, metaphysics, ethics, political theory, aesthetics, psychology. Overlaps especially with politics and philosophy, conceptual/theoretical approach towards the question is more towards philosophy.
ephemera-practical	Almanack, calendars, advertisements, notifications, pamphlet-sized education and science.
history	Historical works, biographies, description of events. Overlaps with literature, as many literary works also take the form of memoir or 'history'. Overlaps with politics, as there is no clear-cut difference between news and recent history.

Table 1. List of the main categories of the classification scheme. It is non-exhaustive, as there are always editions that do not exactly fall into any categories.

Subcategories provide additional information about the publication in the form of keywords that specify some of its aspects in more detail. There can be multiple subcategories per record and they are separated by semicolons. For example, ‘poetry’ and ‘commerce’ are keywords appearing in the data. In total, 216,000 editions have at least one subcategory. Subcategories are not strictly related to any specific main category, but they often have a strong affiliation to certain main categories. For example, there are 22,000 editions with the subcategory poem in the main category ephemera-entertainment and 11,000 in the main category literature and arts. This makes sense, because by default poems go to these categories. However, there are also 810 editions with this subcategory in the main category religion; this also makes sense in the eighteenth-century context, in which religious texts with varying amounts of poetic elements were abundant. Table 2 lists examples of subcategories, and the complete list is provided in the appendix.

Sub Category	**Typical For Main Category:**
act, bill	law
theatre program, almanack	ephemera-practical
petition	politics
sermon, prayer book	religion
metaphysics, logic	philosophy
biography, ancient history	history
agriculture, medicine	science
play, poem	literature and arts

broadside poem, chapbook	ephemera-entertainment

Table 2. Examples of subcategories and their typical relations to main categories.

Supercategories aggregate main categories together to form even more extensive groups. One application is that,by using supercategories, one can avoid problems caused by systematic convolutions between main categories. For example, if there is no reason to distinguish political pamphleteering from legislative documents, it can be useful to treat them as one category of jurisprudence, which is not affected by the heavy tendency of politics to ‘bleed’ to the category of law at the level of the main category. Supercategories are listed in table 3.

Supercategory	**Merges the Main Categories of:**
jurisprudence	law, politics
practical	science, ephemera-practical, education
leisure	literature and arts, ephemera-entertainment
philosophy	philosophy
history	history
religion	religion

Table 3. Supercategories of the classification scheme.

4. Classification Process

The data set was created by mapping the ESTC records to main categories and subcategories (described below) in six steps. An edition that acquired a label from one step was not passed to the next one. All steps were looped via the work-id. This means that if one edition of a work acquired a classification during any of the given steps, then it was projected to other editions of that work as well. This means that all editions of the same work should have the same label.

First, all periodicals detected based on microfilm and bibliographic (e.g. in which bibliographies is the record being cited) data in the ESTC were labelled as periodicals. That is to say, they were given the main category ‘literature and arts’ and the subcategory ‘periodical’.

Second, all manually annotated works (e.g. works with at least one labelled edition) were given a main category and, if available, subcategories based on the manually annotated data. This resulted in more than 100,000 editions with a main category. Many (30,000) of the manually annotated documents also received a subcategory.

Third, a hand-selected set of ESTC keywords about topic, genre and format of editions were mapped to corresponding main categories and subcategories, and these mappings were used to label works. In exploring potentially useful topic and genre keywords, statistical approaches were used as heuristic tools, but all keywords were manually checked. Some keywords were also mapped to a more specific subcategory (e.g. the keyword ‘medicine’ to the subcategory ‘medicine’ in addition to the main category ‘science’).

Fourth, a similar mapping-to-categories approach was used with ngrams (from monograms to pentagrams) of full ESTC titles. For example, ngrams of the form ‘sermon preached at’ or ‘proclamations of X’ are expressions that tell us the type of a publication (religion and law) very accurately. Here too, statistical tools were only used as heuristical finding aids; the final approval of a ngram to the list of ngram-to -category-mappings was a human-made choice. Some ngrams also mapped to more specific subcategories (e.g. ‘sermon preached at X’ to the subcategory ‘sermon’).

Fifth, ESTC keywords that were deemed ‘not good enough but potentially useful’ in the third step were used in a similar manner. For example, the keyword ‘England’ was used at this stage, as it often maps to political pamphlets. However, ‘England’ itself can relate to other than political topics as well without great leaps of imagination, so the keyword was deemed to carry greater risks than keywords like ‘sermon’ used in the third step, which describe the edition very accurately. That is to say, the fifth step consists of riskier keywords used as a last resort, if something more precise is not available.

Sixth, ngrams that were deemed ‘not good enough but potentially useful’ in the fourth step were used in a similar manner as in step five.

After the evaluation, main categories were mapped to supercategories that aggregated them further.

In steps three, four, five and six, it was possible for conflicting labellings to emerge. For example, an edition could have both the keywords ‘sermon’ (religion) and ‘political controversy’ (politics). In these instances, the category of an edition was decided based on the prominence (**prominence score**) of keywords or ngrams related to different main categories (the more prominent the better) and by the number of topics and ngrams of the edition mapping to the category (the more the better). We describe this step in exact terms in the appendix. Each classified edition of a work then ‘votes’ for the main category of the entire work, with its vote being equal to the prominence score. Works classified based on keywords inherit all subcategories to which the keywords of the work map and works classified based on title inherit all subcategories to which the ngrams of the title map.

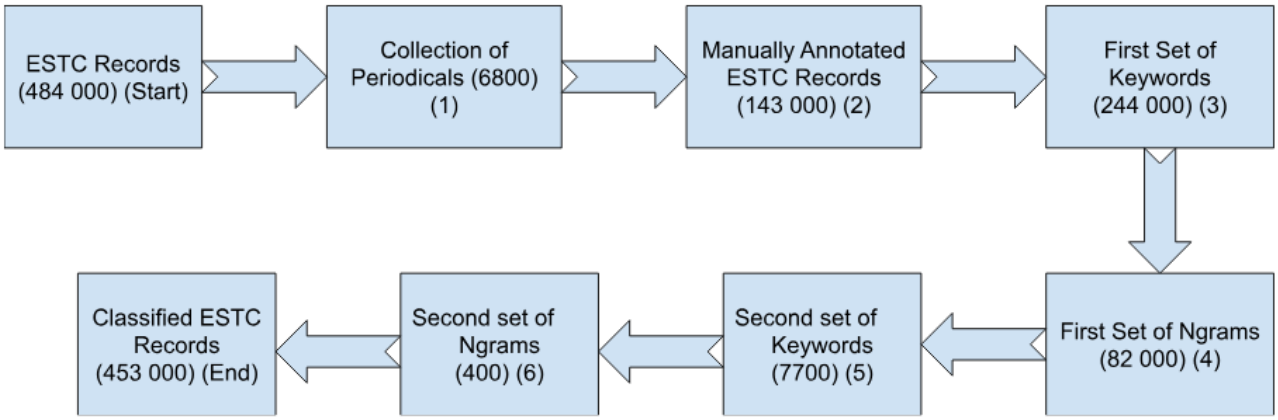

Figure 1. The workflow to categorise records in the ESTC. This part of the workflow is followed by evaluation of the classifications.

5. Evaluation

Overview

The quality of the main categories has been evaluated in a semi-quantitative way. Two researchers (authors) went through a sample of 1000 records with main categories and 500 without them (e.g. the process did not result in a main category for the record). The sample contained information about the size of the document (book, pamphlet or in-between), the suggested main category (and if available, the suggested subcategory), as well as the full title. We decided to use the full title, as eighteenth-century works often contained lengthy subtitles that acted like a table of contents and therefore added important details about the nature of a work.

Both annotators scored labels of the sample of 1000 records with the following scheme: 0 (wrong main category), 0.5 (plausible, but not the best choice) or 1 (best or one among equally good choices). 0 was applied when there was no good argument for assigning the category for a record. For example, a prayer book classified as science would have received a 0. The classification 0.5 (plausible) was used if there was a good argument for the classification, but the researcher felt that the argument was weaker than for some other category or categories. For example, a pamphlet about religious toleration could have obtained the main category religion in the workflow. This would not be a wrong classification per se, but (many would argue) politics would be even more appropriate, if the pamphlet was more focused on political rights related to religion rather than to religious topics as such. In this case, 0.5 would be an appropriate score. 1 was used if the category of the record was the best one available. Record could receive this score even if there was an equally good (but not better) alternative. For example, a comical poem about commercial policy could go to either politics or literature and arts equally well. For records that obtained a score of 0.5 or lower, the researchers also suggested the best possible label. For the records with missing labels, the researchers provided

their best suggestion for a label. If the metadata, such as size and title, did not give any indications about the genre category, both annotators consulted the original text and discussed pertinent issues with each other. These strategies helped to reduce the number of records for which it was not possible to evaluate the assigned label or give an updated label. In the end, missing evaluation score was a rare phenomenon, occurring only for 0.8% of our 1000-record sample. For the 500-record sample of missing labels, we were unable to assign an updated label to 6.2% of records.

This manually evaluated data was used to estimate the precision of the main categories (how often they are right) as well as their coverage (how well a main category covers the relevant records). Coverage also considers the estimated effect of unclassified and mislabelled editions. In calculating coverage, a classification was considered to be right if the sum score of the two annotators was ‘good’ (at least 1.5/2). A ‘good’ score can be thought of as the label receiving mostly favourable evaluations (at least one score of 1 and one score of 0.5). Plausible refers to a sum score that was at least 1 and best to a 2/2 sum score. The distribution of genres among editions that did not receive a label from the workflow was calculated by averaging the suggested labels of the two annotators.[11] Averaging was also done to the labels suggested by the annotators for those editions that failed to get a ‘good’ score.

Score (IT)	Score (KH)	N
1	1	803
0.5	0.5	91
0	0	44
1	0.5	34
0.5	0	5
0.5	1	5
1	0	1

Table 4. Editions from the evaluated sample that obtained a score from both annotators, aggregated by the score they received. N=992.

The estimates of precision and coverage are summarised in table 5. For all categories, 88-100 percent of the assigned editions are at least plausible. Instances where one annotator gave a score of 0 (wholly incorrect category) and the other a score of 1 (best category) were vanishingly rare. Whenever this kind of scoring did occur, both annotators subsequently consulted the original text and afterwards

[11] Averaging was used to solve instances in which the two annotators suggested a different main category. There were 34 (7% of the sample) of such instances.

discussed the reason for their respective scoring. This discussion often revealed that one of the annotators had consulted the original text before assigning the score, whereas the other had given their score based solely on the title. These measures helped to flag and resolve extreme inconsistencies and acted as additional verification steps for edge cases. The final distribution of scores is presented in table 4. There is more variation in terms of the best category, as it fluctuates between 58-92 percent, with most categories having a precision greater than 80 percent and all except one greater than 75 percent. The coverage of all categories is estimated to be between 57-97 percent. All categories mostly include editions that they should and cover most of the relevant records.

Main Category	Precision (plausible)	Precision (good)	Precision (best)	Coverage
education	0.96	0.87	0.83	0.57
ephemera-entertainment	0.97	0.93	0.86	0.97
ephemera-practical	0.96	0.92	0.88	0.76
history	0.88	0.67	0.58	0.63
law	0.93	0.82	0.78	0.9
Literature and arts	0.96	0.8	0.78	0.82
philosophy	1	1	0.92	0.81
politics	0.94	0.8	0.77	0.58
religion	0.95	0.89	0.82	0.89
science	1	0.9	0.88	0.75

Table 5. The evaluation results of the main categories.

In some categories, most or all ‘errors’ are in fact ambiguities, as extension of the definition of right category from ‘the best one possible’ to ‘plausible’ erases most of them. This is also true at an aggregate level: 81 percent of the editions in the sample obtained the best possible score, whereas 94 percent were at least plausible. This means that almost a fifth of the assigned categories were at least somewhat contested by at least one annotator, but more than two thirds of these contested labels were still considered to be reasonable (plausible) even if not ideal. These 0.5-point divergences often reflect the two annotators’ slightly different views on whether a category label could be considered the best-fitting (and thus receive a score of 1) or whether they regarded another category as even more suitable, therefore assigning a 0.5. Overall, both annotators were of the opinion that this in-between score was a valuable addition to the scoring system, enabling a more finely grained evaluation of the precision of a category and accounting for the complexity of genre in the Early Modern period.

The conceptualisation of a single main category being ‘right’ was meant to be a simplification that consciously ignores many of the complexities of Early Modern books. Nevertheless, we implemented tentative analyses on how often the notion of one category that fit the editions better than the others ‘failed’. This helped us to understand how much information was lost with the simplification and how much working within its confines affected our calculations. 28 editions in the sample received a score of 1 from both annotators and were commented during the evaluation process to have equally good candidates existing. Together with other indicators of disagreement about the best label (1-0.5, 1-0 score pairs from the annotators), we found 68 editions out of 1000 that could go equally well to several main categories. 34 of the 500 editions from the sample without editions received different main categories from the two annotators. This too, is 7 percent of the sample, and possibly somewhat inflated by the fact that the annotators discussed divergences less than with the other sample. Both approaches to measuring probably also miss some editions that are problematic for a single-category schema, but the proportion should rise considerably from 7% threaten the assumption that a single main category works well enough in most instances to be a useful way to think and compute[12] about genre.

Distant and Close Reading of Correlated Categories

Our aim is not to downplay the errors in the data. In fact, they are rather interesting, and demonstrate the problematic aspects of any attempted classification scheme of Early Modern books and pamphlets. These overlaps also illustrate the differing literary conventions and traditions of the eighteenth century, where genres considered distinct in the twenty-first century were combined much more frequently and loosely. As Figure 2 demonstrates, the errors are very unequally distributed. Some pairs of real and assigned categories barely exist in the evaluated sample, but others are very significant. Especially, tens of percents of documents best classified as politics end up in law. History ‘bleeds’ into politics, literature and arts, and religion to a considerable degree in Figure 2, and Figure 3 illustrates how it mistakenly includes many editions better classified as literature, politics, education or science. In contrast, categories related to formalised and well-established discourses like law, science and religion are more resilient against losing records to other categories.

[12] For example, averaging the numbers of editions belonging to different genres when the two annotators were disagreeing is a concession to the fact that the conceptualisation of a single main category per edition was not always applicable when doing calculations.

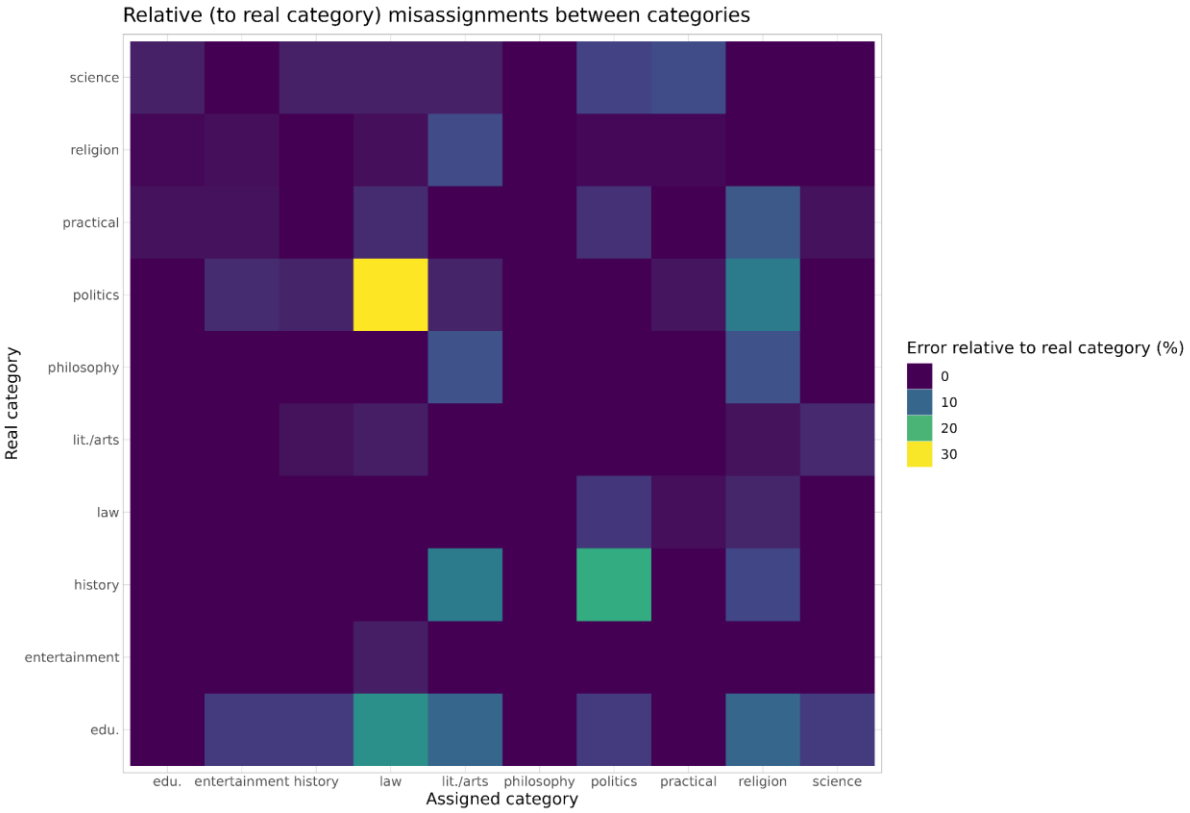

Figure 2. Errors in the evaluated sample compared to the size of the real category.

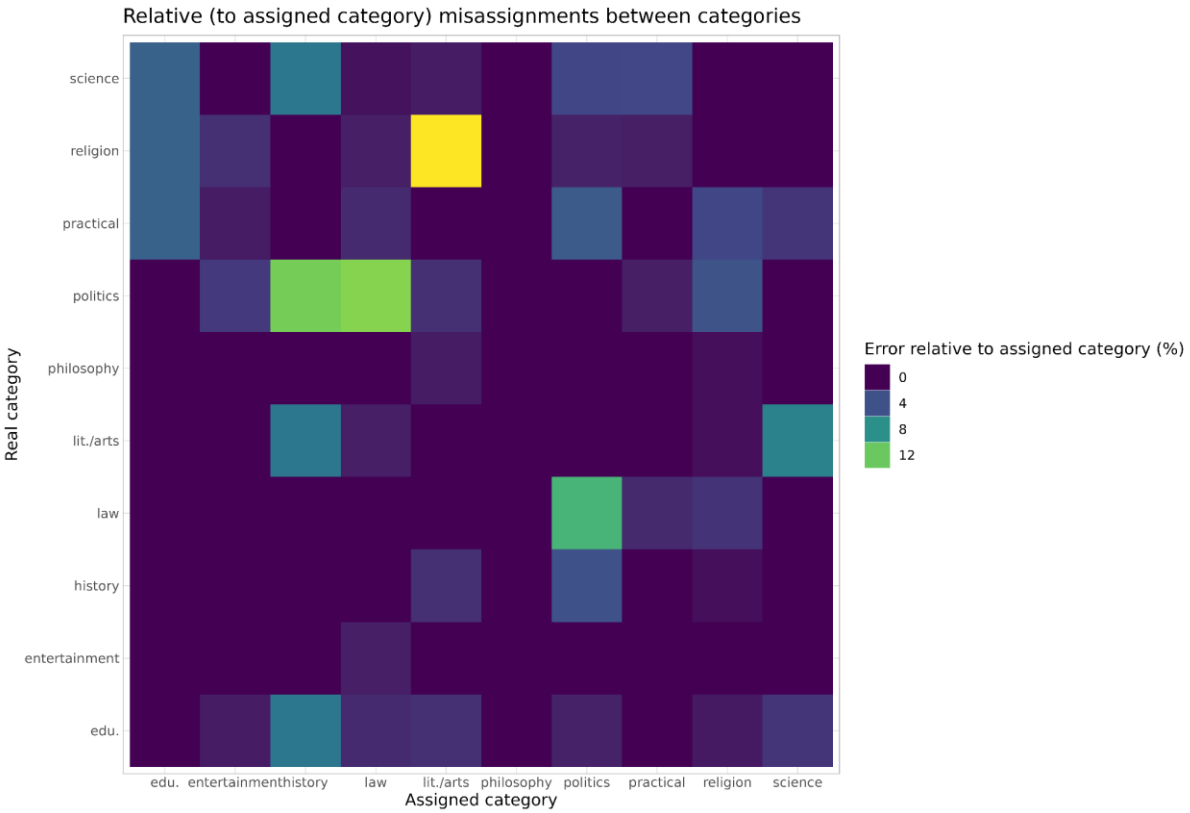

Figure 3. Errors in the evaluated sample compared to the size of the assigned category.

The results of the quantitative error analysis prompted us to take additional steps. We created the supercategories that combine main categories together. The reasoning was that, in some instances, it is better to treat the heavily correlated categories (especially politics and law) as one category. What these supercategories lose in granularity, they get in robustness. The second step was to examine some of the problematic categories more closely. Our conclusion is that the problems of categorisation largely reflect problems of applying clear-cut modern categories to Early Modern books, but also provide tentative directions for further research on classifying the genre and topic of ESTC records.

In the close reading, we especially focused on the connection between law and politics, the overlap between religion and politics, as well as the issue of history, the category with the worst precision. The connection between politics and law is rather straightforward: politics is often about law and distinguished less by substance than register or rhetoric. The following is the title of a record (T52822) from our sample classified as law:

The worth of liberty consider'd: in a letter to a Member of the House of Commons upon the question, how far the late act against the immoderate use of spiritous liquors may affect the properties of all the people.

Both researchers gave this label a score of 0.5, as the document's content is an argumentative discourse about legislation rather than primary legislation as such. However, it certainly relates to legislation. This thematic overlap does not mean that the difference between the two categories would be somehow arbitrary: it is plausible that the argumentative language of Early Modern politics (or even its description like the one given in the title) could be distinguished from language strictly confined to legislative issues as such. For making such distinctions, state-of-the-art tools of machine learning or computational linguistics could be useful.

Another common overlap occurs between religion and politics, which shows the complexity of applying modern genre categories to historical periods. In the Early Modern period, religion and politics were much more intertwined than is the case today. Sermons in particular were often very topical, especially during the English Civil War, and could touch on important political issues of the day. Debates about the rights of religious minorities–such as Dissenters (Protestants who were not members of the state Church of England) and Jews–contained a combination of religious and political arguments. These prevalent overlaps led the annotators to reflect on and discuss how to deal with cases where the outward form suggests one genre (here: religion) and the topic, as indicated by the title, is at least a mix of two genres (here: politics and religion) or leaning more towards a genre that is different to the form (here: politics).

Below is a record (T202404) from our sample with the assigned category of religion, illustrating the overlap between religion and politics and the difficulty of evaluating which of the two genres is the most fitting description for such works.

The young cobler of Glocester: or, Magna Charta [sic] discours'd of between a poor man and his wife. As also several high-church principles discussed, ... Together, with reflections upon several of the vices of the high-church clergy, by the cobler and his wife.

One of the annotators assigned a score of 1 on the basis that the subtitle ‘reflections upon several of the vices of the high-church clergy’ indicates that the work is primarily religious in nature, and thus the most fitting genre label. The other annotator agreed that religion was a plausible genre, but thought the work was even more likely to have a political thrust, as evidenced by the title’s references to Magna Carta and High Church principles (which often went hand in hand with support for the Tory party). Here, the annotators’ motivations for assigning their respective scores are a testament to how their respective judgement is shaped by their own academic interest and domain knowledge.

The category with the lowest precision, history, provides interesting insights into how history writing was practised in the Early Modern period. On the face of it, a title containing phrases such as ‘a history of’ should serve as a reasonably good predictor that the work in question belongs to the genre of history. However, during our annotation process, we discovered that works seemingly historical in nature could, in fact, sometimes be entirely fictional. Distinguishing between biographies of real and invented persons was particularly difficult as Early Modern biographers tended to title their works ‘true’ or ‘authentic’ regardless of whether their subject ever existed, and fiction writers frequently mimicked the writing style and conventions of nonfiction genres (e.g. ESTC records T110215, T173534). An additional challenge was deciding on the cutoff point at which a work is considered historical. In the Early Modern period, many historiographers wrote works covering events up to the present, often with a view to making an intervention in contemporaneous political debates. Deciding whether a work describing recent events, such as a war that happened 20 years ago, belongs to the genre of history or politics was not always straightforward, especially since the title itself often gives little insight into whether the work is primarily focused on describing the past or using the past for present-day politics. One example of this overlap is record N9975 below:

M[emoirs and observations] of the occurrences of Europe, Since the treaties of Nimeguen and Ryswick, with relation to the present treaty at Utrecht. Shewing, that it is of the last importance that England have a footing upon the continent.

Its initial label was history, but both annotators gave a score of 0.5 and suggested that politics was the more fitting label. The period covered by the work, as indicated by the title, is 1678 (Treaty of Nijmegen) until 1712, its year of publication. When the

work was published, negotiations for the Treaty of Utrecht were still ongoing and the subtitle clearly indicates the author's contemporaneous political concerns.

The dividing line between religion and history—rather distinct categories in the twenty-first century—was also much blurrier in the Early Modern period. Biblical criticism as an academic sub-field was still in its infancy and writers of religious history freely mixed real events occurring in biblical times (e.g. the siege of Jerusalem in 70 CE) and events and figures only described in the Bible and for which historical veracity cannot be established (e.g. ESTC record W21041). Assigning such works to either history or religion was often a judgement call for us annotators since it was difficult to estimate the balance between primarily religious elements and primarily historical elements without consulting the work in question. All these complexities explain why the precision for history is significantly lower than is the case for all other categories. We sought to mitigate these difficulties in multiple ways, e.g. discussing particularly complex cases, assigning a score of 0.5, and looking at a few original texts to see if a general rule for annotating such cases could be derived.

6. Applications of the Data, Reflections and Future Work

The new category data introduced in this paper significantly extends the possibilities of using the ESTC and its subsets in quantitative analyses of Early Modern history. As established in Section 2, the ESTC (alongside its subsets) has been the primary resource in attempts to chart the topic or genre composition of the publishing outputs of Early Modern Britain. For the eighteenth century - which comprises most of the editions in the ESTC - this has only been possible for subsets or samples of the whole catalogue until now. Furthermore, there has been no framework that would have brought the entire catalogue spanning more than 300 years within the confines of a single concise and comprehensive classification scheme. Our categorisation covers 94 percent of all records in the ESTC. To our best knowledge, Figure 4 is the most comprehensive depiction of the composition of Early Modern book publishing in the English-speaking world yet produced.

Initial examination of Figure 4 adds significant support to claims made about large-scale trends of book publishing in pre-existing studies and points out to less 'expected' developments that should be studied in more detail in the future. We can see the same relative decline of religion and the rise of literature during the eighteenth century as noted by earlier and more limited quantitative studies (Suarez 2009; Tolonen et al. 2021). Our results make it increasingly likely that these developments are robust and do not reduce to particulars of any specific classification scheme or subset of ESTC editions analysed.

A development that warrants closer scrutiny is how the proportion of politics jumps up during the decade of the English Civil War (1640s) and remains at a higher level

than before ever since. As the relationship between the Civil War and the emergence of a ‘public sphere’ of political discourse in England has been discussed in qualitative research (Zaret 2000), the finding is potentially valuable if it is reasonably accurate.[13] However, as the Civil War decade is also a threshold that separates different segments of the ESTC (consisting of several preceding catalogues), transitions happening in the data in the 1640s should be analysed with especial care, as they can also be linked to different biases and classification conventions of the different segments (Tiihonen 2020).

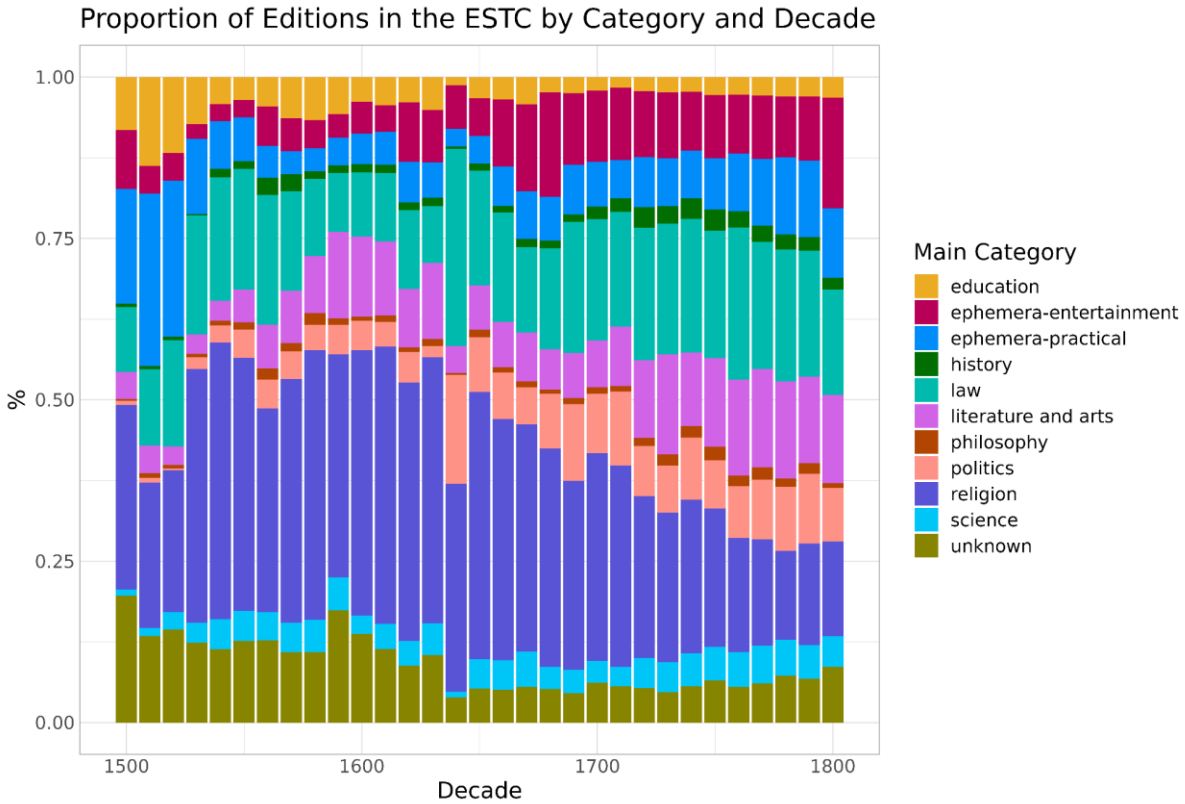

Figure 4. Main Categories of editions in the ESTC 1500-1800.

The categorisation also helps to study other data sets. The ESTC is the default resource used to check for topic or genre-related biases in other collections of Early Modern books like ECCO (Tolonen, Mäkelä, and Lahti 2022) or the Early Modern subset of HathiTrust Digital Library (Almelhem et al. 2023). A more comprehensive classification of the ESTC enables researchers making use of other data sets to compare the genre or topic composition of their data set to it. This can reveal biases in the data set that is being studied.

[13] The fact that politics was a major topic in the publications of the 1640s is in itself an extremely well-known and much-studied phenomenon (Raymond 2003). However, here we can compare the decade of the Civil War to roughly 150 years both before and after it at the scale of the entire ESTC.

In Figure 5, we compare the category composition of the ESTC to ECCO, which is the most comprehensive full-text (e.g. also covers the contents of an edition and not only its metadata) collection of books printed in Britain in the eighteenth century. As the figure demonstrates, the collection is not an accurate representation of all the surviving editions in the ESTC. Legislation and practical ephemera are underrepresented, whereas science, history and philosophy are overrepresented to a considerable degree. This comparison extends preceding quantitative work on the biases of ECCO (Tolonen, Mäkelä, and Lahti 2022) that could not utilise a fully classified ESTC. We can now argue more confidently that ECCO is especially biased to include texts related to 'Enlightenment categories' such as philosophy and science and that it has a tendency to exclude administrative and practical documents, although they comprise a significant part of Early Modern publications.

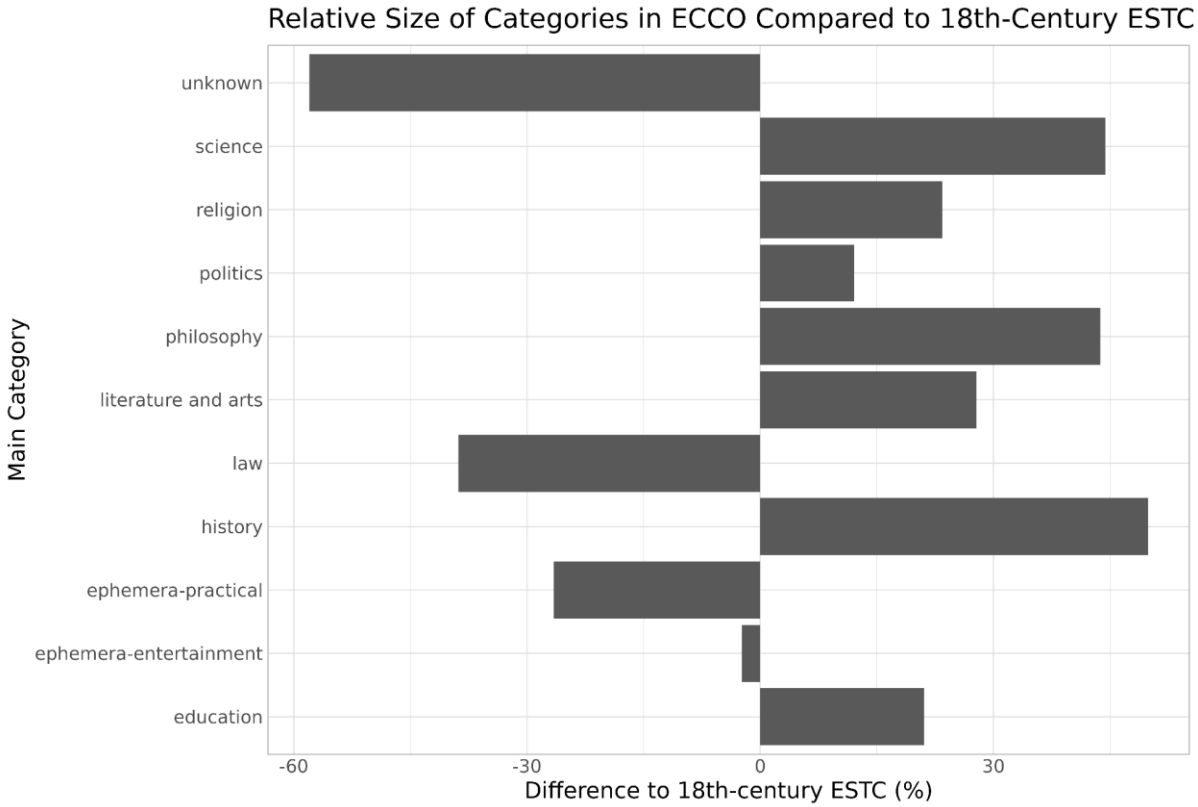

Figure 5. The relative 'bias' of ECCO compared to the ESTC by main category.

More important than the immediate findings are the yet unexplored possibilities offered by the new data and the lessons learned from the classification process. Jointly, our workflow and evaluation demonstrate that a good understanding of the data can enable a simple, eclectic and transparent workflow that produces desired results even for a complicated and challenging data set like the ESTC. A systematic, but hands-on and humanities-oriented, evaluation process of the results helped us to pinpoint more precisely to which degree the data enrichment workflow was working: the manual evaluation enabled us to consider the plausibility of an assigned category

in addition to whether it was ‘right’, and this ended up being a very fruitful way to approach the often complicated Early Modern texts.

Our evaluation revealed that certain document types (e.g. those in politics and law) in the ESTC might benefit from a more complex classifier utilising tools from machine learning or computational linguistics, as keywords and titles can miss the essential elements relevant for their classification. Despite their problems discussed earlier, unsupervised and semi-supervised approaches might have their use in discovering ‘genres’ that are characterised by a complex configuration of linguistic, topical or even visual cues.[14] Thus, our results highlight the necessity of projects that approach Early Modern genres and related questions from various angles, and the possibility that new techniques like multimodal learning might help to derive more holistic computational interpretations of what Early Modern genres were.

The process described here will be updated. A step to be taken in the near future is to enrich the pre-1700 ESTC with the keyword information from the USTC. This should significantly extend the coverage of records with a label before the eighteenth century. Another potential step is to create a version of the data in which multiple categories can exist for any given document. This would not be technically difficult to implement and, as our evaluation results demonstrate, ambiguity of categories was an important element of Early Modern books published in the Anglosphere, making the case for a multi-label version strong.

Funding sources

We thank the Finnish Cultural Foundation and the Research Council of Finland (grants no. 347709 and 333716) for funding the work on this article.

Authorship Statement

Iiro Tiihonen: Conceptualisation (lead), Data curation (equal), Formal analysis, Investigation (equal), Methodology, Software, Visualisation, Writing – original draft (equal), Writing – reviewing & editing (equal). **Kira Hinderks:** Conceptualisation (supporting), Data curation (equal), Investigation (equal), Writing – original draft (equal), Writing – reviewing & editing (equal).

This authorship statement uses the CRediT NISO standard.[15]

[14] For example, satirical illustrations could be a distinguishing feature between law and politics in Early Modern publications.
[15] https://www.niso.org/publications/z39104-2022-credit

Acknowledgements

This article is part of the ongoing and future-focused efforts of the Helsinki Computational History Group (COMHIS) to harmonise, standardise and enrich the ESTC. This work builds on previous initiatives within the group and lays the groundwork for continued advancements in the field.

Bibliography

Almelhem, Ali, Murat Iyigun, Austin Kennedy, and Jared Rubin. 2023. 'Enlightenment Ideals and Belief in Progress in the Run-Up to the Industrial Revolution: A Textual Analysis'.

Amory, Hugh. 1983. Review of *Review of Eighteenth-Century British Books: An Author Union Catalogue; Eighteenth-Century British Books: An Index to the Foreign and Provincial Imprints in the Author Union Catalogue, F. J. G. Robinson*, by F. J. G. Robinson, G. Averley, D. R. Esslemont, P. J. Wallis, J. M. Robinson, and C. Wadham. *The Papers of the Bibliographical Society of America* 77 (1). [University of Chicago Press, Bibliographical Society of America]: 80–84.

Baker, K.M. 1964. 'The Early History of the Term "Social Science"'. *Annals of Science* 20 (3): 211–26. doi:10.1080/00033796400203074.

Barr, Bernard. 1981. 'Paying for the Eighteenth Century: The ECBB'. *Library Review* 30 (4): 229–32. doi:10.1108/eb012729.

Corns, Thomas. 1986. 'Publication and Politics, 1640–1661: An SPSS-Based Account of the Thomason Collection of Civil War Tracts'. *Literary and Linguistic Computing* 1 (2): 74–84. doi:10.1093/llc/1.2.74.

Feather, John. 1986. 'British Publishing in the Eighteenth Century: A Preliminary Subject Analysis'. *The Library* s6-VIII (1): 32–46. doi:10.1093/library/s6-VIII.1.32.

Fielding, David, and Shef Rogers. 2017. 'Copyright Payments in Eighteenth-Century Britain, 1701- 1800'. *The Library* 18 (1506): 3–44.

Gants, David. 2002. 'A Quantitative Analysis of the London Book Trade 1614-1618'. *Studies in Bibliography* 55 (January): 185–213.

Gillings, Mathew, and Andrew Hardie. 2023. 'The Interpretation of Topic Models for Scholarly Analysis: An Evaluation and Critique of Current Practice'. *Digital Scholarship in the Humanities* 38 (2): 530–43. doi:10.1093/llc/fqac075.

Hill, Alexandra. 2018. *Lost Books and Printing in London, 1557-1640. An Analysis of the Stationers' Company Registers*. Brill.

Ijaz, Ali, Hege Roivainen, and Leo Lahti. 2019. 'Analytical Edition Detection In Bibliographic Metadata'. In . DataverseNL. doi:10.34894/IWQ5BO.

Jenner, Mark. 2011. 'London'. In *The Oxford History of Popular Print Culture*, edited by Joad Raymond and Gary Kelly, 294–307. Oxford University Press.

Kapoor, Sayash, and Arvind Narayanan. 2023. 'Leakage and the Reproducibility Crisis in Machine-Learning-Based Science'. *Patterns* 4 (9). Elsevier. doi:10.1016/j.patter.2023.100804.

Lahti, Leo, Jani Marjanen, Hege Roivainen, and Mikko Tolonen. 2019. 'Bibliographic Data Science and the History of the Book (c. 1500–1800)'. *Cataloging & Classification Quarterly* 57 (1). Routledge: 5–23. doi:10.1080/01639374.2018.1543747.

McKenzie, Donald. 2002. ‘Printing and Publishing 1557–1700: Constraints on the London Book Trades’. In *The Cambridge History of the Book in Britain*, edited by John Barnard and Donald McKenzie, 4:553–67. The Cambridge History of the Book in Britain. Cambridge University Press. doi:10.1017/CHOL9780521661829.028.

Murphy, Sean. 2019. ‘Shakespeare and His Contemporaries: Designing a Genre Classification Scheme for Early English Books Online 1560–1640’. *ICAME Journal* 43 (1): 59–82. doi:10.2478/icame-2019-0003.

Nicoll, Allardyce. 1959. *History of English Drama 1660–1900: Volume 6, A Short-Title Alphabetical Catalogue of Plays*. Cambridge University Press.

Raymond, Joad. 2003. *Pamphlets and Pamphleteering in Early Modern Britain*. Cambridge University Press.

———. 2011. ‘The Development of the Book Trade In Britain’. In *The Oxford History of Popular Print Culture*, edited by Joad Raymond and Gary Kelly, 59–75. Oxford University Press.

Ryan, Yann Ciarán, and Mikko Tolonen. 2024. ‘The Evolution of Scottish Enlightenment Publishing’. *The Historical Journal* 67 (2). Cambridge University Press: 1–33. doi:10.1017/S0018246X23000614.

Shadrova, Anna. 2021. ‘Topic Models Do Not Model Topics: Epistemological Remarks and Steps towards Best Practices’. *Journal of Data Mining & Digital Humanities* 2021 (October). Episciences.org. doi:10.46298/jdmdh.7595.

Simmons, Richard. 2002. ‘The Cambridge History of the Book in Britain’. In , edited by John Barnard and Donald McKenzie, 4:504–13. Cambridge University Press.

Snyder, Henry L. 1983. Review of *Review of Eighteenth-Century British Books: An Author Union Catalogue Extracted From the British Museum General Catalogue of Printed Books, the Catalogues of the Bodleian Library, and of the University Library, Cambridge.*, by F. J. G. Robinson, G. Averley, D. R. Esslemont, and P. J. Wallis. *Eighteenth-Century Studies* 16 (3). [Johns Hopkins University Press, American Society for Eighteenth-Century Studies (ASECS)]: 342–46. doi:10.2307/2738357.

Suarez, Michael. 2009. ‘Towards a Bibliometric Analysis of the Surviving Record, 1701–1800’. In *The Cambridge History of the Book in Britain: Volume 5: 1695–1830*, edited by Michael Suarez and Michael L. Turner, 5:37–65. The Cambridge History of the Book in Britain. Cambridge: Cambridge University Press. doi:10.1017/CHOL9780521810173.003.

Tiihonen, Iiro. 2020. ‘From Explosion to Implosion. A Quantitative Analysis of the English Civil War Print Production’. Master’s thesis, University of Helsinki.

Tiihonen, Iiro, Leo Lahti, and Mikko Tolonen. 2024. ‘Print Culture and Economic Constraints: A Quantitative Analysis of Book Prices in Eighteenth-Century Britain’. *Explorations in Economic History* 94 (October): 101614. doi:10.1016/j.eeh.2024.101614.

Tolonen, Mikko, Mark J. Hill, Ali Zeeshan Ijaz, Ville Vaara, and Leo Lahti. 2021. ‘Examining the Early Modern Canon: The English Short Title Catalogue and Large-Scale Patterns of Cultural Production’. In *Data Visualization in Enlightenment Literature and Culture*, edited by Ileana Baird, 63–119. Cham: Springer International Publishing. doi:10.1007/978-3-030-54913-8_3.

Tolonen, Mikko, Eetu Mäkelä, and Leo Lahti. 2022. ‘The Anatomy of Eighteenth Century Collections Online (ECCO)’. *Eighteenth-Century Studies* 56 (1). Johns Hopkins University Press: 95–123. doi:10.1353/ecs.2022.0060.

Watt, Tessa. 1990. ‘Publisher, Pedlar, Pot-Poet: The Changing Character of the Broadside Trade, 1550-1640’. In *Spreading the Word. The Distribution Networks of Print 1550-1850*, edited by Robin Myers and Michael Harris, 61–82. Saint Paul’s Bibliographies.

Whitten, David. 1978. ‘Democracy Returns to the Library: The Goldsmiths’-Kress Library of Economic Literature’. *Journal of Economic Literature* 16 (3). American Economic Association: 1004–6.

Zaret, David. 2000. *Origins of Democratic Culture: Printing, Petitions and the Public Sphere in Early-Modern England*. Princeton University Press.

Zhang, Jinbin, Yann Ciarán Ryan, Iiro Rastas, Filip Ginter, Mikko Tolonen, and Rohit Babbar. 2022. ‘Detecting Sequential Genre Change in Eighteenth-Century Texts’. In *Proceedings of the Computational Humanities Research Conference 2022*, edited by Folgert Karsdorp, Alie Lassche, and Kristoffer Nielbo, 3290:243–55. Antwerp, Belgium: CEUR. https://ceur-ws.org/Vol-3290/#short_paper2630.

Appendix

Supplementary Data

act	adventure	advertisement	advice	aesthetics
agriculture	almanack	ancient history	antiquarianism	applied knowledge
architecture	astrology	astronomy	bible	bill
biography	calendar	catalogue	catechism	chemistry
classics	comedy	commerce	conduct of life	cooking
declaration	devotional exercise	dictionary	doctrine	fable
for children	geography and nature	handbook	hobbies and games	human mind
hymn	language	linguistics	logic	manners
mathematics	medicine	memoir	metaphysics	meteorology
military applications	moral	music	navigation	news
novel	opera	painting	pastoral letter	periodical
petition	physics	play	poem	political thought
prayer book	proceeding	proclamation	psalm	report
review	satire	satire poem	sermon	song
speech	technology	theatre program	theoretical	theory of knowledge
travel	trial	utopia		

Appendix table 1. Full list of subcategories

Technical Details

Prominence score F of a main category for an edition **x** and main category **c** was calculated in the following way: **F(c,x) =** $\sum_{k \in (K_x \cap K_c)} ln(Q(k))$

where **k** is a keyword or title ngram in the ESTC, K_x is the set of keywords or title ngrams related to the edition **x** and K_c is the set of keywords or title ngrams related to main category **c.** Function **Q** measures the number of times that the keyword or ngram appears in the ESTC. In a verbal form: the prominence score is a sum of natural logarithms of the prominences of keywords or ngrams that relate to an edition and one main category. Constructed in this manner, the equation favours the information provided by the most common keywords and title ngrams. The main category of the edition was chosen by choosing the category **c** with the highest prominence score.