# Computing Book Parts with EEBO-TCP

Misson, James; Singh, Devani Mandira

**How to cite**

MISSON, James, SINGH, Devani Mandira. Computing Book Parts with EEBO-TCP. In: Book history, 2022, vol. 25, n° 2, p. 503–529.

This publication URL: https://archive-ouverte.unige.ch/unige:165508

# Computing Book Parts with EEBO-TCP

*James Misson and Devani Singh*

## 1. Anatomical Bibliography and EEBO-TCP

Quantitative and computational bibliography have traditionally used variations on "the book" as their base unit of measurement. When a claim such as "145 books were published in Ireland in 1703" is made,[1] a single data point is defined by the edges of the physical object that it counts. This is somewhat at odds with recent trends elsewhere in bibliography. Since Gérard Genette theorized the paratext, historians of the book have increasingly anatomized the objects of their study and paid more attention to books' separate components as discrete entities.[2] The approach is typified in *Book Parts*, an essay collection edited by Dennis Duncan and Adam Smyth which opens the book and breaks it down into its "constituent material components." We can conceive of the book, they write, "not as a stable object, but as a coming together or an alignment of separate component pieces, each possessed of particular conventions and histories"—page numbers, frontispieces, dedications, blurbs, and more.[3] The early handpress period, an era of substantial formal experimentation within the history of the book, has been especially receptive to scholarly approaches which "crumble the wholeness of the book" (as Duncan and Smyth put it).[4] Nor is the pursuit limited to the academic world. Keith Houston's *The Book* is a self-referential object that offers a granular guide to itself.[5] In addition to these overviews, in-depth studies have been written on specific book parts, including footnotes, printed marginalia, and indexes,[6] in the same vein as Margaret M. Smith's classic work *The Title-Page*.[7]

Our impulse to anatomize is not merely an imposition of modern taxonomies onto historical practice. Writers and printers in the early modern period thought of books in parts just as scholars today do. By 1622, for instance, paratextual adornment had become so conventional that "To set

*Book History*, vol. 25, no. 2

forth a booke without an Epistle, were like to the old English proverbe, a blew coat without a badge," according to the stationer who wrote *Othello*'s preface.[8] New title pages were attached to unsold books to give them a second life,[9] and Tiffany Stern has shown that title pages were also detachable, posted around the urban landscape to advertise an edition.[10] Apparatuses like glossaries were likewise an optional appendage. A glossary is added to a 1553 edition of *Pierce the ploughmans crede* "For to occupie this leaffe which els shuld have ben vacant."[11] In other words, the early modern book was a professedly modular object, its discrete units having diverse sources and destinations that are belied by the limited but convenient ontology that "the book" represents.

It's no surprise that quantitative bibliography operates on a different scale, as most of its datasets are drawn from catalogues, originally designed and structured as finding aids and descriptions of editions or individual books.[12] Though we can query these datasets at the level of the books they represent, getting inside these books is harder; such information rarely accommodates questions like "what genres of text were dedications attached to?" or "when did encomia start appearing in English books?" The extraordinary work of the EEBO Text Creation Partnership, however, now allows researchers to open up these books without sacrificing their ability to be counted as data. In this article, we identify and assess a method of uniting the tools of quantitative bibliography with the scale of "anatomical" bibliography—of computationally studying book parts using a resource that is little known but full of potential: EEBO-TCP's division (or *div*) types.[13]

Early modernists have long been familiar with EEBO-TCP. Since 2000, it has been producing high quality transcriptions of British books printed between 1473 and 1700.[14] Most researchers access these texts via ProQuest's EEBO platform, where they are accompanied by the images from which the transcriptions were made. Historical linguists have also analysed the raw files behind these pages in bulk as linguistic corpora (some examples are discussed below), making these transcriptions critical to our modern knowledge of the history of the English language. But the same raw files also contain data that is invaluable to English book history, much of which is suppressed on EEBO's main platform; the files not only transcribe, but describe too, using XML (Extensible Markup Language) to "replicate the structure of the book."[15] The XML markup, which is now freely accessible online, imparts an additional layer of bibliographical information to the EEBO-TCP corpus of transcribed texts.[16] The division types that we study in this article are one feature of this description. In essence, division types are labels like *preface*, *poem*, or *table of contents*, which are assigned to

different parts of the transcribed text, thereby offering an almost entirely untapped seam of descriptive bibliographical metadata. If book historians can exploit this metadata, then it may yield results that are as consequential to the history of the English book as the transcriptions are to the history of the English language.

The layer of description represented by the division types is mostly hidden to general users of EEBO, and we have found only one published study (by John R. Ladd, discussed below) that takes advantage of them to quantitative ends. In order to encourage their further use, this article therefore gives an introduction to EEBO-TCP's division types and provides some demonstrations of how they may be used, based on our own experiments. We also offer caveats. The division types are an imperfect dataset, often unwieldy and, strictly speaking, not designed for the uses we discuss here. Their limits must be fully understood for their potential to be realised. The quantitative use that we propose is therefore in the spirit of what Ryan Cordell has recently termed "speculative bibliography," which "seeks to identify meaningful patterns for exploration within collections that are often messy and unevenly described."[17] We base these demonstrations and caveats on our experience as book historians handling division types to create a database of early modern prefatory paratexts, and we describe our own methods at the end of the article. Our study of division types has been greatly aided by EEBO-TCP's own website,[18] which is a vital source of information for anyone working with the files. It makes accessible a rich archive of documentation and internal correspondence that is key to understanding the dataset and should be lauded as an exemplar of transparency. Given the field's reliance on EEBO-TCP, researchers who wish to use it responsibly should understand and be able to interpret this wealth of documentation. Such critical engagement may extend to the production of independent appraisals of the dataset, as others have done at earlier stages of the EEBO-TCP project.[19] We intend this article to be another contribution to the field's understanding of EEBO-TCP and one that may enable book historians to use it to its full potential. In this spirit of better evaluating and understanding the dataset, our discussion draws attention to the fact that the EEBO-TCP types are the result of collaborative work between international teams of academic editors and professional keyers hired by several offshore data conversion providers: Apex CoVantage, SPi, Aptara, and AELData (companies whose names may be familiar to readers, as they also routinely typeset English-language academic publications). By drawing attention to the potential of division types in this article, we therefore intend to make visible and acknowledge the labour behind the creation of EEBO-TCP.

## 2. EEBO-TCP, Remediation, and Representation

Before discussing the division types themselves, it must be established that they are a product of the same long sequence of remediation that informs EEBO-TCP's texts, and therefore have the same limits. EEBO-TCP's files are representations of the images found in *Early English Books Online* (EEBO); these images are digitisations of UMI/ProQuest's *Early English Books* microfilm series (EEB)—two collections whose history is by now familiar thanks to articles written by Diana Kichuk, Ian Gadd, and Bonnie Mak,[20] which draw from the biography of UMI's founder, Eugene B. Power.[21] Power saw the potential for microfilm technology to make historical texts accessible, and so after founding University Microfilms Incorporated in 1938, he started photographing books listed in the STC and Wing catalogues and sold the films to the research libraries of American universities.[22] By 1988 UMI (now called ProQuest) had completed the bulk of the microfilm series, named *Early English Books* (or EEB), and in 1998 they made it available online in an early form of EEBO: an archive of bi-tonal scans of the microfilms, accompanied by bibliographical data from the recently created English Short Title Catalogue.[23]

EEBO contains images of over 146,000 titles, and over 17 million pages.[24] It is these images that have been used by EEBO-TCP to create marked-up transcriptions of early modern books. This monumental undertaking began in 2000 and continued until 2020, led by teams based primarily at the University of Michigan and the University of Oxford, who worked under the direction of Paul Schaffner and in collaboration with the offshore data conversion providers Apex CoVantage, SPi, Aptara, and AELData.[25] Having reached the end of its second phase, EEBO-TCP is currently "better described as 'in hiatus' than, strictly speaking, 'done'," according to its website, and the files that it has produced are, since August 2020, freely available to anyone and for any purpose. Keyers were hired to transcribe the text from EEBO images, adding XML markup including (among other metadata) the division types discussed in this article. The resulting texts are of impeccable quality. EEBO-TCP's quality control required 99.995% accuracy in tested samples before files were accepted[26]—a rate well beyond even the most advanced OCR software, which still struggles with early modern text and especially the varying quality of EEBO's images that is a consequence of their chain of remediation.

It is no exaggeration to say that EEBO-TCP has revolutionized early modern English studies. Literary critics, for instance, may easily find their

topic of research through keyword searches, even in non-canonical texts; researchers at the Oxford English Dictionary routinely use EEBO-TCP to antedate the earliest known usage of words; linguists may access early modern language untouched by later editors.[27] Besides such word-level studies, however, researchers are increasingly using EEBO-TCP for larger scale projects, drawing conclusions from quantitative or computational results—a trend that is sure to increase since the 2020 data release. The *Early Print* project, for instance, has begun the process of making this data suitable for linguistic analysis by adding linguistic metadata to every word in most of the corpus. This accommodates tools like their n-gram viewer, which can show the relative frequency of a word across time.[28] Similar techniques have been used by Anupam Basu and Joseph Loewenstein to investigate the alleged archaism of Spenser's orthography with statistical methods, demonstrating that the longstanding belief that Spenser archaized his orthography is an exaggeration.[29]

Given its ubiquity, it is surprising that there has not yet been a detailed independent study of the extent to which EEBO-TCP's remediations have delimited the corpus. Mark Algee-Hewitt et al. productively invoke the simile of the Russian Matryoshka doll to understand the relationship between historical text corpora and historical reality. "The corpus is thus smaller than the archive, which is smaller than the published," they write, "like three Russian dolls, fitting neatly into one another."[30] EEBO-TCP is no different, and this doll has many layers: printed books, surviving books, books included in the STC and Wing catalogues, books selected or available for microfilming, microfilms scanned to digital images, and digital images selected for transcription. Our corpus gets smaller at each of these stages, so if we are to have any confidence in the results of quantitative studies then we must hope that the innermost doll at least resembles the outermost in its proportions if not size. Some of these stages have been accounted for. Alexandra Hill's research on survival suggests that as many as 45% of books from this period may have been lost (though this might best be taken as a worst-case scenario).[31] Historical curation can skew these rates of survival—witnessed most famously in the huge spike of extant books from 1640, attributable to the civil war pamphlets collected by George Thomason in the seventeenth century and preserved by the British Library as the Thomason Tracts.[32] How many of these surviving books make it onto EEBO ? Its website claims to represent "almost every work printed," but not every edition printed, and certainly not every variant—an "illusion of comprehensiveness" that Ian Gadd has discussed.[33] Furthermore, within each work, pages are occasionally missing or otherwise illegible.

Beyond the account given on its website, no equivalent study of the biases inherent to EEBO-TCP currently exists. As the website says, EEBO-TCP covers "approximately 50% of the texts featured" on EEBO,[34] but the question of *which* 50% is difficult to answer owing to the variety of factors that influenced inclusion. The first phase of the project favored authors mentioned in the *New Cambridge Bibliography of English Literature*; books could also be nominated by participating libraries, and even a degree of "more or less random selection" was involved. Coverage of works was prioritised over editions—EEBO-TCP is therefore generally unsuitable for measuring differences between texts of works. And, by the second phase, English text was favored (though many Latin works appear and all Welsh works, so long as they appear in the STC).[35] Until a more detailed inventory is published, any computational use of EEBO-TCP must be mindful of the potential for distortion inherent to its selection. Such an understanding would allow book historians to harness the opportunities offered by EEBO-TCP's unprecedented scale and accuracy, and to do so responsibly.

## 3. EEBO-TCP's Markup

An EEBO-TCP file affords not only an accurate transcription of a book's text but also combines this transcription with descriptive metadata in the form of XML that can be exploited to identify book parts in the files. As carefully created as the texts themselves, this markup makes non-textual aspects of early modern books analysable on a computational scale, but this potential remains largely undeveloped. Anupam Basu has previously noticed this lacuna, observing that "Most large-scale analysis sees the text as a purely linguistic construct," a "bag-of-words," rather than something that has structured, formal properties.[36] Basu performs an experiment that is an extremely effective demonstration of the possibilities inherent to this markup: he analyses EEBO-TCP files purely by their distribution of tags, regardless of the text they contain, and algorithmically identifies books that are formally similar.[37] Remarkably, Basu's algorithm finds not only books of the same genre but even of the same author, based on their formal structure alone. The top match for Thomas Middleton's play *Michaelmas Term*, for instance, is another of Middleton's plays.[38] But despite this clear demonstration of its value, EEBO-TCP's markup is generally suppressed. General users access these files in a mediated form, either through the JISC *Historical Texts* platform or, more likely, through ProQuest's *Early English Books*

*Online*, where some (but not all) of the markup is used for basic formatting and navigation within a book. Those wishing to access the full markup must do some further digging.[39]

This suppression does no justice to the wealth of data found in these files, which contain the full work of the EEBO-TCP keyers and editors: texts tagged with XML according to a schema adapted from the Text Encoding Initiative (TEI) guidelines to accommodate the characteristics of early print.[40] Detailed definitions of the broader application of XML are available, but when applied to books, XML can encode non-verbal characteristics of textual content in ways that are both human- and machine-readable. Each text is split into parts (called *elements*) which form the building blocks of an XML file. Elements are demarcated by an opening tag and a closing tag that describe the structure of the text on a general level; for instance, headings, paragraphs, and general divisions of the text are all tagged. An element can enclose other elements, replicating the hierarchy of the book. The structure of a book is thereby described via its parts, and a typical EEBO-TCP structure might look, in a very simplified form, like that in figure 1. In line twenty-one, for example, a heading in the text is opened with the tag <*HEAD*> and in line twenty-three, it is closed with </*HEAD*>. As is seen in the example, non-specific textual divisions are enclosed within <*DIV*> tags. The range of information captured by EEBO-TCP's XML in this fashion is impressive: speaker headings in dramatic texts, marginalia, and lists are just some of the book parts tagged that appeal to the research interests of book historians. A complete list of the features tagged in EEBO-TCP can be found on the taggers' "cheat sheet".[41] These tags are placed within a higher-level structure that is standardized across files. Most EEBO-TCP files have at least a *body* element, which contains the main text of the work or the main content of the book. Many also have a *front* and a *back* element (as in figure 1), which contain the front and back matter of the book. Within the *front*, *body*, and *back* are divisions (also called *divs*)—a general-purpose tag assigned to elements that are discrete units of text, as determined by the keyer.[42] These divisions are numbered to denote their depth. A *div2* element, for instance, is enclosed within a *div1* element, as with the encomia texts in figure 1. On one level, the encomia are grouped together as a separate division in the book (*div1*, lines 9–16) but within this, each individual encomium has been given a further *div2* tag (lines 10–12, 13–15), supplying it with an additional level of detail and making it discoverable as a separate text.

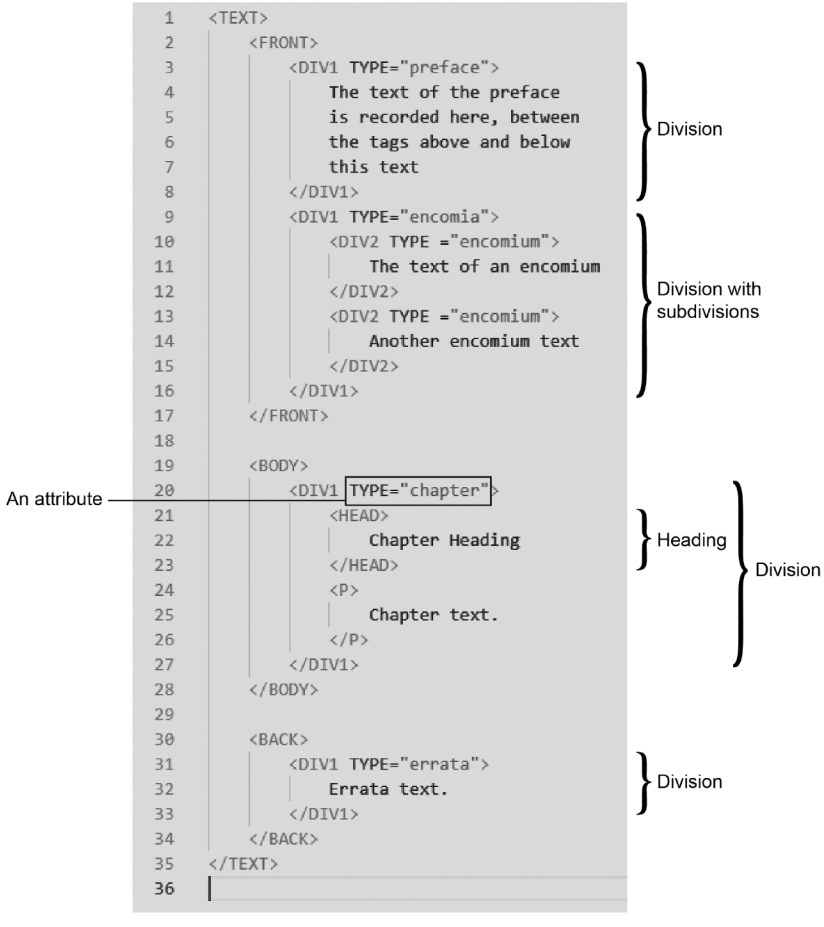


**Figure 1.** A simplified EEBO-TCP XML file.

Readers will have noticed that some of the divisions in figure 1 contain more specific information than the name of the tag. XML contains a feature called an *attribute*, an optional addition to a tag. Attribute names can be defined by whoever designs the XML schema, and in line twenty (and elsewhere) we see an example of EEBO-TCP's *type* attribute, where a division has been given a label, in this case, *chapter*. Remarkably, such labels, or division types, have been assigned to every one of the 1,479,565 divisions in the corpus.[43] The relationships between the different components of relevance to our discussion—text, divisions, attributes, and types—are visually depicted in figure 2.

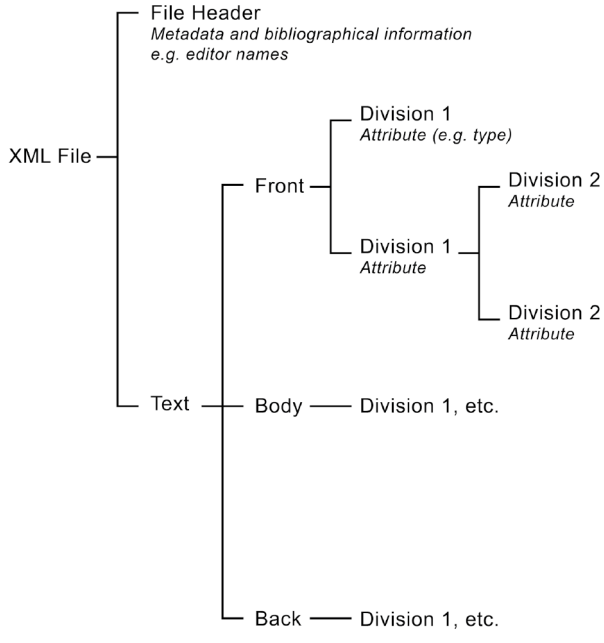


Figure 2. The structure of different elements within an EEBO-TCP file.

## 4. EEBO-TCP's Division Types

Where generic division tags demarcate the parts of an early modern book, the type attribute assigned to a division is used to offer more specificity about the kind of text the division contains. This exhaustive description of EEBO-TCP's divisions is a remarkable feat of bibliography that has great, untapped potential. The division types range from certain formal or bibliographical descriptions (such as *sonnet* or *imprimatur*), to generic terms (like *encomium*), and even to descriptions of the text's content (*list of shipwreck victims*). Across the EEBO-TCP corpus, there are 11,790 unique division types. In our own experience, contextualizing types within the *front*, *body*, or *back* element has been a useful way of further increasing their specificity. Tables 2, 3, and 4, therefore show the top twenty most frequently used division types in each of these sections.[44]

**Table 2. The twenty most frequent division types in front sections of the EEBO-TCP corpus.**

| *Front division type* | *Total* |
|---|---|
| title page | 46,157 |
| dedication | 15,520 |
| to the reader | 11,331 |
| table of contents | 5,617 |
| part | 5,048 |
| preface | 4,681 |
| encomium | 3,137 |
| errata | 2,840 |
| license | 2,516 |
| poem | 2,269 |
| chapter | 2,017 |
| frontispiece | 1,687 |
| imprimatur | 1,602 |
| section | 1,233 |
| illustration | 1,165 |
| dramatis personae | 853 |
| prologue | 838 |
| letter | 806 |
| half title | 804 |
| publisher's advertisement | 697 |

**Table 3. The twenty most frequent division types in body sections of the EEBO-TCP corpus.**

| *Body division type* | *Total* |
|---|---|
| chapter | 212,205 |
| part | 188,538 |
| section | 173,130 |
| poem | 37,159 |
| text | 28,576 |
| entry | 27,976 |
| letter | 26,647 |
| subpart | 23,550 |
| subsection | 23,361 |
| recipe | 20,236 |
| question | 19,924 |
| verse | 18,968 |
| prayer | 17,045 |
| psalm | 16,521 |
| sermon | 14,904 |
| song | 13,668 |
| epigram | 13,535 |
| scene | 13,308 |
| article | 12,099 |
| book | 10,835 |

**Table 4. The twenty most frequent division types in back sections of the EEBO-TCP corpus.**

| *Back division type* | *Total* |
|---|---|
| colophon | 12,941 |
| part | 9,195 |
| section | 3,897 |
| errata | 3,858 |
| publisher's advertisement | 3,043 |
| table of contents | 1,445 |
| index | 1,197 |
| license | 993 |
| postscript | 697 |
| chapter | 656 |
| letter | 590 |
| imprimatur | 536 |
| subpart | 525 |
| book | 496 |
| to the reader | 458 |
| poem | 442 |
| epilogue | 415 |
| appendix | 414 |
| document | 363 |
| advertisement | 298 |

The creation of the division type corpus was a collaborative enterprise. Work was split between professional keyers and EEBO-TCP's editors, and was supervised by Paul Schaffner, the project's manager. The keyers were first asked to assign a type based on the following six rules, whose significance is such that we reproduce them verbatim and in full :[45]

1. Use the designation supplied by the book itself. "Chapter 3" should be recorded as <DIV1 TYPE="chapter">
2. Use lower-case throughout ("chapter" not "Chapter").
3. If the designation is not in English, and there is a ready equivalent in English, use the English. E.g., for "pars" or "partie" use "part"; for "capitulum" or "chapitre" or "cm." or "chapt." or "cap." use "chapter".

   If the designation in the book is a verbose version of a common English term, use the simpler form. E.g., if the book says "Prefatory Remarks by the Author," you shouldn't be afraid to translate this into <DIV1 TYPE="preface">

Otherwise, use whatever is there.

4. If there is no designation in the book, and the <DIV> is used to mark a series of items of similar type, use a term describing the form or genre shared by the items. E.g., in a book of poems, use

   <DIV1 TYPE="poem">
   <DIV1 TYPE="poem">

   See further under Poetry, below.

5. If there is no designation in the book, and the <DIV> is used to mark a series of items of dissimilar type, or if there is no series at all, just use a term that describes the form of the item as generically as can be (<DIV1 TYPE="letter">; <DIV1 TYPE="preface">)

6. If none of these rules apply, do not supply any value for the TYPE attribute.

In those cases when rule six was followed and the keyer did not assign a type, these gaps were filled by the editors at a later stage in the workflow. Editors also reviewed the types supplied by the keyers, changing them if necessary. The types assigned by keyers and editors were then subject to cursory review by Schaffner.[46] The workflow for the assigning of types was thus designed with two layers of editorial oversight.

As these rules suggest, wherever possible the choice of division type was prompted by the text contained in the book, often found in the printed headings relevant to that division. We also find many types that are drawn from the running headings of the book, though running headings themselves are not recorded in EEBO-TCP. But many division types have been assigned in the absence of these sources, drawing on information external to the book, and what is apparent from EEBO-TCP's rules is that assignation of division types could sometimes require historical knowledge. Rule three, for example, requires the translation of early modern Latin or English into modern English, and keyers were likewise asked to extract relevant information from verbose early modern headings. When no clear prompt was extant in the book, they applied bibliographical knowledge to interpret the text, using taxonomies of genre or form, as required in rules four and five. We see, therefore, that although the keyers were not early modern specialists, the assigning of division types required many of the skills normally

associated with textual editing. Besides letter-by-letter transcription, keyers were also applying historical and bibliographical knowledge to the analysis of early modern content and its context, the results of which were supplemented by editors.

## 5. Uses of Division Types

In the early stages of EEBO-TCP, division types seemed to be created for a use that did not yet exist. An internal document that sought to clarify their function states that division types are "primarily useful for navigation in a book" but also that this primary use "is to some extent potential rather than real because much of the information is suppressed in the current interface."[47] As of 2020, some of this potential has been realised. One of the many improvements made by ProQuest's new EEBO interface uses division types for the kind of navigation imagined in that document. If a user opens the EEBO page for the 1630 edition of John Taylor's complete *Workes*, for instance, they are provided a table of contents in a sidebar that allows them to navigate between textual divisions.[48] Most of the headings in this table are drawn from the heading tags in the EEBO-TCP file, such as the first "encomium" division, headed "To the Author, Iohn Taylor". But for those divisions without printed headings, the division type is used, thereby listing the book's "Title page," "Encomia," and "Dedication," etcetera in EEBO's navigational sidebar.

Another use is suggested in the same document. Division types are "secondarily useful as a means of searching or of limiting searches." That is, keyword searches may be performed on EEBO-TCP texts targeting only certain parts of books. This function was available in EEBO's former Chadwyck-Healey interface and continues in that of ProQuest via "field codes" that may be appended to keyword searches. The thirty field codes offered allow a small selection of common division types to be targeted for searching, such as *colophon* (*CPN*) and *to the reader* (*TTR*).[49] EEBO-TCP, however, is conscious of the deficiency of this method: due to a "lack of control on the vocabulary used for types," we cannot be sure that we are searching every desired text .[50] A search for a keyword in *TTR*, for example, would limit the search to divisions that have been assigned the *to the reader* type, but not divisions with the type *to non-English-speaking readers*. Despite the lack of controlled language, this remains, for many division types, a reliable, if limited, way for the average user to find keywords in a certain book part.

The same principle, however, can be employed to do work on a much larger scale, as John R. Ladd has recently demonstrated. Ladd's study exploits division types to reveal the networks that are distributed across early modern dedications.[51] By drawing on names extracted from texts in EEBO-TCP with the *dedication* division type,[52] Ladd delimits his data to individuals appearing in a dedicatory context.

As Ladd describes, his use of division types is a method of locating relevant texts (in this case, dedications) on a large scale before analysing their content. But we can also use division types in ways that do not target text for extraction but, like Basu's experiment, analyse book parts or structures regardless of the text they may contain, thereby shifting the use of EEBO-TCP from literary or linguistic history towards book history. A study of a specific book part will here illustrate the tremendous potential of the division types for our knowledge of the early modern book. Encomia are the short texts that often preceded an early modern work to commend the author or another individual. EEBO-TCP's division type *encomium* (and its variants, such as *introductory encomium*) provide a means of tracing the big picture of the encomium's role in English book history. We began with the simple question of frequency over time: roughly when did encomia start appearing in English books, and how common is their appearance? Such a picture could be extracted from EEBO-TCP by counting the number of files from each year that contain divisions of the type *encomium*, visualized in chart 1.[53] The chart shows that, after a couple of false starts around 1530 and 1545, a trend for encomia rapidly takes off from 1562, spiking in 1576 when they appear in 14% of EEBO-TCP files. Here, the unit of measurement is "a book that contains one or more encomium"; an alternative approach would be to count the total number of encomium divisions irrespective of the files they appear in, and so our unit becomes the individual book part—the encomium itself. The results (chart 2) again show a trend beginning in 1562 that peaks in 1576 (thirty-four encomia), before dropping suddenly to just two in 1585. This technique can also be used to identify structurally anomalous books. A significant spike in encomia in 1611, for instance, is caused by the parodic encomia of Thomas Coryat's *Coryat's Crudities* and a pirated anthology of these published in the same year, *The Odcombian Banquet*.

With the range of EEBO-TCP's tagging, similar pictures could be generated not just for encomia, but for epigraphs, acrostics, lists, emblems, and dozens of other parts of the early printed English book. To carry these studies to greater depth, the quantification performed above would be combined

with the qualitative analysis associated with the use of division types for locating relevant texts. Features of encomia such as their length, language, authors, and addressees, could be mapped over time for the entire corpus. By introducing other variables into the foundation outlined here—like a book's format, genre, author, or stationers involved in its making—this study could be developed to shed light on the presence and popularity of encomia in certain segments of the book trade. Likewise, book parts do not exist in isolation, and similar frequency counts for comparable division types could be combined with the results given above to determine whether encomia tend to appear alone or cluster with other paratexts, thereby permitting a rich and nuanced overview of the appearance of encomia across a large corpus of early modern printed books.

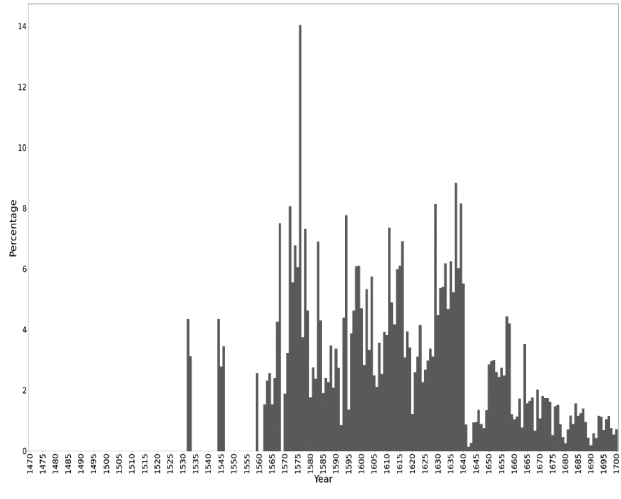


**Chart 1.** Percentage of EEBO-TCP files containing divisions with type *encomia* by year.

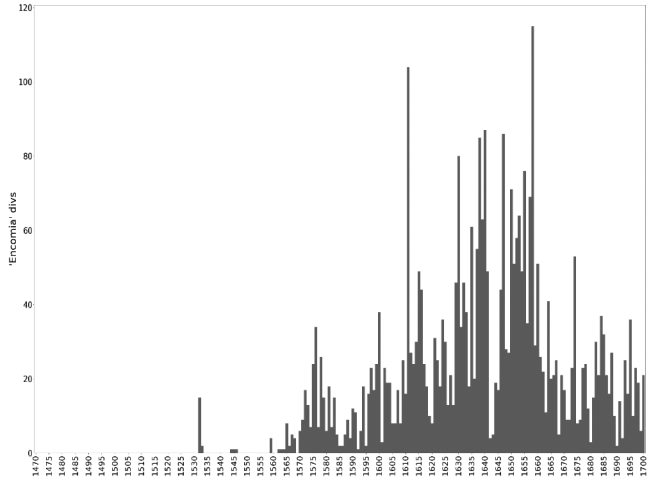


**Chart 2.** Divisions with type *encomia* by year.

## 6. The Problem of Standardization

The use of division types must be governed by an awareness of the unstandardized state of the dataset, bearing in mind EEBO-TCP's own caveat that there is a "lack of control on the vocabulary used for types." It is this lack that has resulted in such wide variety. In total, there are 11,790 unique division types in the EEBO-TCP corpus. Furthermore, the frequency distribution of these 11,790 types is skewed, with many more infrequently used types than there are frequently used types. This distribution can be gauged in two different ways. First, we can examine how frequently a type appears in the whole corpus—as has been done for tables 2, 3, and 4, showing the twenty most used types in the front, body, and back sections. At the other end of this distribution are the types that are rarely used. In fact, most of the division types are used only once—we find 7,267 single-use division types out of 11,790 (62% of the total), and 90% of division types are used fewer than 22 times across the whole corpus. Low-use types are typically descriptive rather than typological (such as *index of Hebrew words*) and/or quasi-transcriptions of headings that result in a variant of a high-use division type (such as *to the sincerely professing reader*).

The second way of gauging the distribution of types lies in the number of individual files in which a division type appears. It is necessary to consider this separately from the frequency of a type, as certain types at first seem

frequent enough to be useful to quantitative studies, but in fact all of their occurrences may appear in only a handful of files. The example of the type *subentry* is illustrative. Appearing 5,651 times in the whole corpus, this type is the 33rd most frequent division type; and yet, all of these instances appear in the XML of a single book—William Patten's 1575 *The calender of Scripture*, an encyclopaedic companion to people and places named in the Bible.[54] In this case, the keyer or editor has assigned a type to each of the book's entries, choosing the idiosyncratic label *subentry*. Other such single-file types that appear many times are likewise idiosyncratic choices of those assigning the type or the unique formal demands of a specific book. Of all the division types, 8,899 types (or 75%) are unique to individual files, whereas six division types appear in more than 10,000 files.[55]

The EEBO-TCP corpus can therefore be thought of as containing at least two kinds of division types: those that are high-use and usually typological, like *table of contents*, *preface*, and *colophon*, and those that are low-use and usually descriptive, like *table of fireworks*, *advert for mouthwash*, and *description of siege tower*. Crucially, many low-use or descriptive division types may be referring to the same thing as a single high-use, typological division type, and projects interested in one kind of book part must therefore account for this if the division types' potential beyond navigation is to be realised. Our study of encomia above, for instance, takes a deliberately naïve approach, counting instances of division type *encomium*, but not its low-use variants such as *acrostic encomium*, or *introductory encomium*; a more accurate study would first consolidate all of the descriptive variants under the typological division type *encomium*. This is easy enough for encomia: the presence of the word *encomium* in the descriptive types allows them to be grouped together easily; more advanced techniques of consolidating types can exploit the fact that modern English has been used for the division types. Their removal of the orthographical variation that normally makes early modern English resistant to computational analysis proves to be one of their great contributions. Natural Language Processing (NLP) could, for instance, be used to parse the division type corpus. Variants of the *to the reader* type, such as *to the Protestant reader*, could be standardized by removing all adjectives; descriptive division types like *catalogue of birds* or *list of ships from Falmouth* could be reduced to a typological type by removing the prepositional phrases. Regular expressions, a programming method which can identify patterns within a text, likewise provide a reliable method of consolidation. For instance, a regular expression can identify text that contains the words *to* and *reader* with any word between them,

thereby capturing all the variants of *to the reader*, including *compositor to the reader*, *to the fun-loving reader*, *to the botanical reader*, and so on. Doing so groups 164 variant division types under a single *to the reader* group: a group that now represents 14,270 texts, adding 1,783 texts to the 12,487 texts with the type *to the reader*.

But the problems posed by variant and low-use types remain when we want to draw comparisons between multiple division types. Clearly something like *list of horse names* is not quantifiable alongside something so fundamental as *title page*. However, the line between a typological and descriptive type is not always so clear-cut, and definitions of low- and high- use will be peculiar to different types and the research questions to which they are being applied. Furthermore, many of the low-use types defy consolidation. Single-use types like *mnemonic* are, in our experience, not an indicator that the keyers or editors have failed to rigorously categorize but an accurate reflection of the dizzying array of parts that could appear in an early modern book—an accuracy that is jeopardized by the pursuit of a universal standard. This means that any responsible standardization of division types is likely to be tailored towards the specific aims of individual projects. There are as many ways of categorizing the types as there are of doing anatomical bibliography. Our approach is described below.

## 7. Division Types and the "To the Reader" Project

Our work with division types was conducted in the context of the ongoing project "To the Reader: The English Preface in Print," which studies the emergence of early modern prefatory paratexts printed until 1640. In line with the principles outlined above, we used the division types in EEBO-TCP's XML markup to assemble a corpus of texts relevant to the project's aims and scope. To determine which division types should be considered for inclusion in our corpus, we first used a Python script to generate a master list of all unique division types (at depths *div1* and *div2*) in the front, body, and back sections for all of the files up to 1640. Compared to the numbers of division types for the entire EEBO-TCP dataset, this was relatively small: a total of 1,276 unique division types in the front section, for instance, and only 640 in the back section. The difference between these numbers and the number of unique division types in the EEBO-TCP dataset as a whole is attributable to the exponential rise in the number of files after 1640. Given

this not unmanageable size, we decided to use the lists generated as a menu, hand-selecting the division types we judged relevant to our project.

This was straightforward for many division types, whose names are generic enough to be self-explanatory: *preface*, for instance, was in, as was *to the reader*. But *table of contents*, *title page*, and *dramatis personae* (for example) were out, because those parts of the book are unlikely to contain addresses to readers, the project's principal interest. Working through the division types without their accompanying texts, however, we found many whose relevance was not so clear to us. What kind of text did the division type *oath* refer to, for instance? We deferred these to a second round of selection. Those texts with ambiguous division types were extracted by the Python script along with the unambiguous ones to create a maximal corpus; we then read the texts with ambiguous types, which allowed us to make a final decision on that type and therefore to adjust the script and run it again. During this second round, we encountered more ambiguity. Sometimes a division type included both texts that were patently of relevance to our project and texts that were not. Did we want to include *errata*, for instance, which is normally used for tables listing errata but sometimes includes a note to the reader apologising for the errors? Such types were identified in the second round as instances which needed to be evaluated on a case-by-case basis. This required that each text be read individually to determine whether or not to include it based on its content rather than its division type.

Our method was therefore much more involved—and therefore more accurate—than the automated selection of just one division type used in the encomia study above. The method was more accurate because it synthesized manual and automated inclusion, and used division types as a foundation that was then supplemented by the reading of each text. Despite this high level of involvement, the synthesis of automated and bespoke selection dramatically reduced the work required to create a corpus. The majority of texts included are products of the first round of division type selection, which automated the bulk of the corpus building before it was fine-tuned. The corpus derived from our three rounds of selection contains 19,140 prefatory paratexts, represented by 593 unique division types. The process of selection in three rounds gave us a deep familiarity with the types contained in our corpus of paratexts, and we were therefore able to manually sort them into a system of twelve categories. Table 5 presents our categories and a sample of the division types contained within them.

**Table 5. Division types sorted into categories.**

| *Category* | *Number of unique division types in category* | *Examples* | *Number of texts in category* |
|---|---|---|---|
| Dedications | 29 | translator's dedication, dedicatory letter | 7,186 |
| To the Reader | 128 | to Catholic readers, editor to the reader | 4,994 |
| Poems | 49 | elegy, dedicatory ode, encomium | 3,153 |
| Prefaces & Introductions | 48 | author's preface, prefatory material, prelude | 1,633 |
| Errata notes | 9 | errata, corrigenda, erratum | 490 |
| Arguments & Summaries | 25 | argument, prose summary, summary of treatise | 301 |
| Letters | 20 | prefatory letter, to the Duke of York, letter to his son | 237 |
| Afterwords | 21 | Caxton's epilogue, editor's envoy, printer's afterword | 156 |
| Bibliographical addresses | 35 | to the translator, author to publisher, printer to book | 146 |
| Prayers | 3 | translator's prayer, verse prayer, hymn and prayer | 91 |
| Other | 226 | exhortation, printer's note, invocation | 772 |

Division types have therefore provided the foundation from which our entire corpus of paratexts could be classified, and by building on the work of EEBO-TCP's keyers and editors we have developed an intuitive system of sorting which would otherwise have been much more labor-intensive. The classifications can now be used in synthesis with other variables to begin to answer the project's core question about the emergence of prefatory paratexts and addresses to readers in printed English books. It can also be used to extract textual data from those items which are of interest. Our grouping of the types offers a more reliable method of locating texts than the use of any one division type alone. Our classified corpus is thus intended both as the basis for a qualitative study of addresses to the reader and as a database to be queried quantitatively.

We are in the process of combining this corpus of prefatory paratexts with bibliographical data from various sources to create a queryable database. By uniting the texts with, for instance, genre and language data[56] as well as the standard bibliographical data provided by the ESTC, we can ask a broad range of questions about our chosen book parts and about the printed book trade more generally. For example, and most fundamentally, how did the trend of adding dedications or prefaces to the reader change over time? Chart 3 shows these figures, decade by decade, for our three largest categories created from the division types: dedications, addresses to the reader, and poems. Each line represents the percentage of EEBO-TCP files in each decade that contain one or more of these types of text. The chart shows that, in the EEBO-TCP corpus, addresses to the reader and dedications both rise in tandem in the 1520s, from appearing in around 0% of books in the 1510s, to 37% and 45% of books respectively in the 1570s. From then, both kinds of book part steadily maintain these rates until the 1610s. Though their upward trends are similar in timing, a boom in the number of dedications in the 1570s (from 27% to 45%) means that nearly half of the books represented by EEBO-TCP contain a dedication between the 1570s and 1610s. Both addresses to the reader and dedications then see a dip going into the 1620s. Prefatory poems likewise see a rise in use in the 1570s, appearing in 15% of books in that decade, but level out at a lower rate of between 7 and 10% for the remainder of the period.

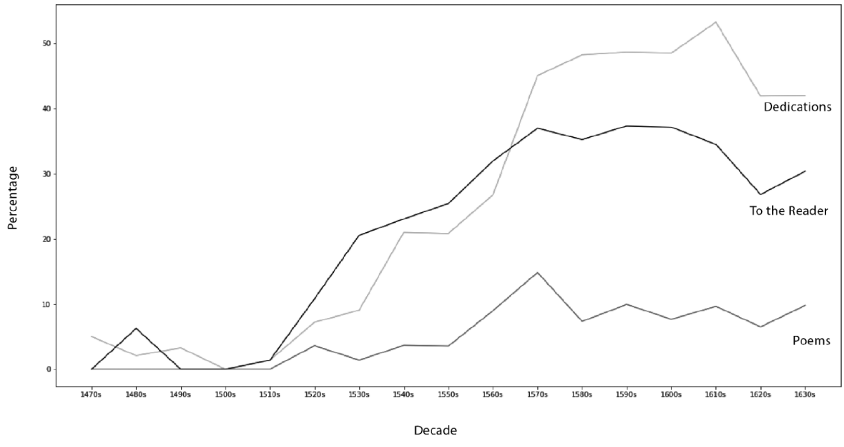


**Chart 3. Percentage of EEBO-TCP files containing the largest three categories of division types.**

On a more fine-grained level, we could ask which genres attracted the most encomia, what kind of prefatory paratext was most likely to be in a different language to the main text, or whether octavos differ from folios in their use of epistles to the reader. That is, the quantitative study of book parts facilitated by EEBO-TCP's division types has enabled us to gather a richer, more ambitious corpus than would otherwise be practical and to ask more nuanced questions of it. Behind such questions is an understanding of the early modern printed book as a locus of historical and socio-cultural meaning. Assessed computationally, the addresses and epistles to readers we have gathered using EEBO-TCP's division types allow us to confirm, contest, or elaborate upon certain *grands récits* within literary and book history. For example, it has been observed (of the patronage of English poetry in the sixteenth and early seventeenth centuries) that "another set of social relations was emerging in which the patron was ultimately eclipsed by the increasing sociocultural authority of authors as well as by the economic and interpretive importance of the reader."[57] Our computational study of printed prefaces allows us to assess the bibliographical record for evidence of such economic and social change and for the supposed rise of the figure of the reader. The approach to division types we have described in this article would allow book historians to undertake similarly detailed and multifaceted studies of other early modern book parts at unprecedented scales, as demonstrated by our article's investigations into encomia. The division types, we have suggested, are an underappreciated but powerful feature of

the EEBO-TCP files and our work within the "To the Reader" project has convinced us of their immense utility to book historical scholarship.

* * *

EEBO-TCP has provided researchers with a dataset of field-changing value. It represents the culmination of two decades of scholarship, coordination, and collaboration between the researchers, librarians, keyers, and editors who contributed to it. The division types discussed in this article are the fruits of their intellectual work, and we wish to conclude by addressing and acknowledging the labor that underpins EEBO-TCP and its markup. As Michael Gavin has pointed out (and as was noted above), EEBO-TCP texts were keyed by data conversion specialists working for "vendors"—a cluster of companies, including Apex CoVantage, SPi, Aptara, and AELData, with offices in India and the Philippines.[58] In a response to Gavin, Peter C. Herman has expressed concern about the source of EEBO-TCP's labor, characterizing Gavin's statement of the fact as an "admission" and "dropping a bomb"—a reaction which seems prompted only by the locations of the keyers, whose workplaces are "not countries known for high wages and worker benefits."[59] This question of the conditions under which the EEBO-TCP work was performed is a legitimate one, but it is one which we believe is best pursued by researchers with the expertise to appropriately contextualise the socio-economic situation of the keyers.[60] In the absence of information about the companies' practices, it is imprudent for early modernists to imply that the EEBO-TCP keying contracted out to offshore companies took place under problematic labour conditions by virtue of their location in particular countries.

We do agree with Herman in his broad opinion that EEBO-TCP would do well to make more information about the keyers available. While their role in the project is no secret (a list of forty-three vendor staff members is published on EEBO-TCP's website),[61] they are at present not credited in the files themselves. While the US- or UK-based editor is often named in the XML header, the keyer remains anonymous—a contravention of the system of acknowledgement and citation otherwise used for scholarly labor, and one which we hope will be duly addressed in future iterations of the EEBO-TCP project. This is not only a question of ethics and decorum, but also a matter of enabling the informed use of the corpus. As Andie Silva argues, "Transparency in the credit and design of digital projects is bound to produce better users and, ideally, better evaluators of digital labour."[62] In

the meantime, we believe that all users of EEBO-TCP should inform themselves as best they can of the magnitude and quality of the keyers' work. Our study of division types is a contribution towards this understanding; we ultimately show that keyers have not only transcribed the texts, but, by assigning division types, also performed the kind of bibliographical and historical analysis that was required of EEBO-TCP's editors. We intend for this article to inform and inspire early modernists and book historians to explore the potential of division types for the study of early printed book parts and hope that, as the computational use of EEBO-TCP expands in the years to come, due attention will also be given to the keyers and the dozens of librarians, students, and information professionals who collaborated to build this extraordinary resource.

## Acknowledgement

This work was supported by the Swiss National Science Foundation/Fonds National Suisse under a grant (no. 179809) held by Devani Singh. We are grateful to Yuri Cowan, Jasmeer Virdee, and the two anonymous readers at *Book History* for their generous comments on earlier versions of this article.

## Notes

1. This example is drawn from Michael F. Suarez, S.J., "Towards a Bibliometric Analysis of the Surviving Record," in *The Cambridge History of the Book in Britain, Volume V*, ed. Michael F. Suarez, S.J. and Michael L. Turner (Cambridge: Cambridge University Press, 2009), 50.

2. Gérard Genette, *Paratexts: Thresholds of Interpretation,* trans. Jane E. Lewin (Cambridge: Cambridge University Press, 1997).

3. Dennis Duncan and Adam Smyth, "Introductions," in *Book Parts*, ed. Dennis Duncan and Adam Smyth (Oxford: Oxford University Press, 2019), 5.

4. The approach is also amply demonstrated by the studies in *Renaissance Paratexts*, ed. Helen Smith and Louise Wilson (Cambridge: Cambridge University Press, 2011).

5. Keith Houston, *The Book: A Cover-to-Cover Exploration of the Most Powerful Object of Our Time* (New York: W. W. Norton & Company, 2016).

6. Anthony Grafton, *The Footnote: A Curious History* (Cambridge: Harvard University Press, 1997); William W. E. Slights, *Managing Readers: Printed Marginalia in English Renaissance Books* (Ann Arbor: University of Michigan Press, 2001); Dennis Duncan, *Index, a History of The*: *A Bookish Adventure from Medieval Manuscripts to the Digital Age* (New York: W.W. Norton, 2022)

7. Margaret M. Smith, *The Title-Page: Its Early Development 1460–1510* (New Castle: Oak Knoll Press, 2001).

8. Thomas Walkley, "The Stationer to the Reader," in *The tragœdy of Othello, the Moore of Venice* (London: Nicholas Okes, 1622), sig. A2$^{r}$.

9. Jonathan R. Olson, "'Newly Amended and Much Enlarged': Claims of Novelty and Enlargement on the Title Pages of Reprints in the Early Modern English Book Trade," *History of European Ideas* 42, no. 5 (2016): 619.

10. Tiffany Stern, "'On each Wall and Corner Poast': Playbills, Title-pages, and Advertising in Early Modern London," in *English Literary Renaissance* 36, no. 1 (2006): 57–89.

11. Anonymous, *Pierce the ploughmans crede* (London: Reynold Wolfe, 1553), STC 19904, sig. D3$^{v}$.

12. The work of Mikko Tolonen and the COMHIS group at the University of Helsinki has, for instance, turned the entries of the *English Short Title Catalogue* into data. See, especially, Mikko Tolonen, Mark J. Hill, Ali Zeeshan Ijaz, Ville Vaara, and Leo Lahti, "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*, ed. Ileano Baird (Basingstoke: Palgrave Macmillan, 2021), 63–119.

13. In general usage, the acronym EEBO-TCP (Early English Books Online Text Creation Partnership) refers both to the organisation behind the corpus, and the corpus itself.

14. Its full purview is that of the Short Title Catalogue (STC) and Wing catalogue: books printed in the British Isles and its colonies, and books printed in English and other British languages between 1473 and 1700. Alfred W. Pollard and Gilbert R. Redgrave, eds., *A Short-Title Catalogue of Books Printed in England, Scotland, & Ireland and of English Books Printed Abroad 1473-1640*, 2nd edn, rev. William A. Jackson, F. S. Ferguson and Katharine F. Pantzer, 3 vols (London: Bibliographical Society, 1976–91); Donald Wing, ed., *Short-Title Catalogue of Books Printed in England, Scotland, Ireland, Wales, and British America, and of English Books Printed in Other Countries, 1641–1700*, 2nd ed. (New York: MLA, 1972–78).

15. Rebecca Welzenbach, "Transcribed by Hand, Owned by Libraries, Made for Everyone: EEBO-TCP in 2012," 8, https://deepblue.lib.umich.edu/bitstream/handle/2027.42/94307/welzenbach-oxfordeebotcp-2012.pdf?sequence=1.

16. On 1 August 2020, the 60,326 files that make up the EEBO-TCP corpus were released for general use.

17. Ryan Cordell, "Speculative Bibliography," *Anglia* 138, no. 3 (2020): 522.

18. https://textcreationpartnership.org.

19. E.g. Brian Vickers, "Is EEBO-TCP/LION suitable for attribution studies?", *Early Modern Literary Studies* 21, (2019): 1–34.

20. Diana Kichuk, "Metamorphosis: Remediation in *Early English Books Online (EEBO)*," *Literary and Linguistic Computing* 22, (2007): 291–303; Ian Gadd, "The Use and Misuse of *Early English Books Online*," *Literature Compass* 6, (2009): 680–692; Bonnie Mak, "Archaeology of a Digitization," *Journal of the Association for Information Science and Technology* 65, (2014): 1515–1526.

21. The following account is condensed from Kichuk, Gadd, and Mak, drawing on Eugene B. Power's autobiography, *Edition of One* (Ann Arbor: University Microfilms Inc., 1990).

22. Power, *Edition of One*, 28–29.

23. The readily available accounts of EEB's, and therefore EEBO's, history are almost entirely derived from Power's autobiography; this historiography deserves some scrutiny, and given that bibliographical projects of this scale are always collaborative, a full history of EEBO beyond Power, utilizing the extensive archives of EEBO-related material held by the University of Michigan, is overdue. Mak has begun this by acknowledging the contribution of Margaret Harwick to the distribution of EEB, as well as photographers Lucia Moholy and Adele Kibre's work on UMI's collaboration with US intelligence agencies during World War II. Mak, "Archaeology," 1518.

24. ProQuest, "About Early English Books Online," https://search.proquest.com/eebo/productfulldescdetail?accountid=14624.

25. The TCP website (https://textcreationpartnership.org/about-the-tcp/tcp-staff/) acknowledges the wide range of people who were "involved in the conception, promotion, and execution" of the project, but for the purpose of understanding the EEBO-TCP markup, we are chiefly interested in the work carried out by the project's keyers and editors. On the website, editors are credited under two separate categories: "primary production staff" and "part-time and students."

26. Welzenbach, "Transcribed by Hand," 3.

27. Louise Mycock and James Misson, "Lone Pronoun Tags in Early Modern English: ProTag Constructions in the Dramas of Jonson, Marlowe and Shakespeare," *English Language & Linguistics* 25, (2021): 379–407.

28. EarlyPrint, "EarlyPrint Library," https://earlyprint.org/lab/.

29. Anupam Basu and Joseph Loewenstein, "Spenser's Spell: Archaism and Historical Stylometrics," *Spenser Studies* 33, (2019): 63–102.

30. Mark Algee-Hewitt, Sarah Allison, Marissa Gemma, Ryan Heuser, Franco Moretti, and Hannah Walser, *Canon/Archive: Large Scale Dynamics in the Literary Field,* Pamphlets of the Stanford Literary Lab, 11 (Stanford: 2018).

31. Alexandra Hill, *Lost Books and Printing in London, 1557–1640: An Analysis of the Stationers' Company Register* (Leiden: Brill, 2018), 3.

32. See Iiro Tiihonen, "From Explosion to Implosion: A Quantitative Analysis of the English Civil War Print Production," (Master's Thesis, University of Helsinki, 2020), https://helda.helsinki.fi/handle/10138/314293.

33. Gadd, "Use and Misuse," 686.

34. ProQuest, "About Early English Books Online," https://search.proquest.com/eebo/productfulldescdetail?accountid=14624.

35. Text Creation Partnership, *About EEBO-TCP*, "How we selected texts," https://textcreationpartnership.org/tcp-texts/eebo-tcp-early-english-books-online/.

36. Anupam Basu, "Form and Computation: A Case Study," in *Digital Milton*, ed. David Currell (Basingstoke: Palgrave Macmillan, 2018): 119.

37. Users may generate their own comparisons through EarlyPrint's "Discovery Engine," https://earlyprint.org/lab/tool_discovery_engine.html.

38. Basu, "Form and Computation," 122.

39. The current corpus of raw XML files, hosted on box.com, is linked from the EEBO-TCP website's FAQs (a location that the included "readme.txt" describes as "not a perfect or permanent solution"). See Text Creation Partnership, "readme.txt", https://umich.app.box.com/s/f3mphvepm20akwloqna2.

40. EEBO-TCP's XML corpus is available in two forms. While the P5 corpus conforms more precisely to the up-to-date TEI guidelines, it is currently incomplete, and so the P4 corpus is "the version that [EEBO-TCP] generally recommend," EEBO-TCP, *FAQs*, "Can I download the raw files?" https://textcreationpartnership.org/faq/. For an account of the conversion of EEBO-TCP P4 to P5, see James Cummings and Sebastian Rahtz, "Kicking and Screaming: Challenges and advantages of bringing TCP texts into line with the Text Encoding Initiative," Bodleian Libraries, University of Oxford (2012).

41. Text Creation Partnership, "EEBO tagging cheat sheet (1)," https://textcreationpartnership.org/docs/dox/cheat.html.

42. For more on DIVs, see *TEI P5: Guidelines for Electronic Text Encoding and Interchange,* "4.1 Divisions of the Body," https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

43. For more on attributes see *TEI P5*, "v. 6 Attributes". Besides *type*, five other attributes have sometimes been assigned to EEBO-TCP's divisions: *rend*, *lang*, *N*, *ID*, and *subtype*. The rarest division attributes are probably not used with enough consistency or widely enough over the EEBO-TCP corpus to be of any use to quantitative methods, but they may offer means by which interesting features can be located. The *rend* (or "rendition") attribute (used 211 times) is occasionally used to describe an unusual orientation or placement of the text that is not possible to render by transcription; the rendition values "invert" and "rotateClockwise," for instance, have been used to describe blocks of sideways and upside-down text which would be of interest to scholars of early modern mise-en-page. This attribute is more commonly used for non-division elements, such as <SEG REND="decorInit">, consistently used to tag decorative initials. *Lang* has been used 11,022 times in the corpus, to record when the language in the division is different to that of the main language of the file. The *N*, or "number," attribute (appearing 609,393 times) accompanies type attributes to enumerate items in an ordered series

of a certain type. So, for instance, a sequence of divisions that are chapters of a book may be tagged as <DIV TYPE="chapter" N="1"> etc. It is as yet unclear how consistently these five least frequent division attributes are used. Researchers interested in quantifying language change in EEBO-TCP books, for instance, would have to assess how much of their corpus is accurately tagged before exploiting the language attribute.

44. The full lists of division types can be found at https://zenodo.org/record/4926442#.Yqyn-OzMJyw (DOI 10.5281/zenodo.4926442).

45. From Text Creation Partnership, *EEBO Text Conversion Project: Keying/Coding specifications,* https://textcreationpartnership.org/docs/dox/instruct.html.

46. We are grateful to Paul Schaffner for his generosity in discussing the project's history with us for the purposes of this article.

47. Text Creation Partnership, "Assigning TYPEs to DIVs," https://textcreationpartnership.org/docs/dox/DIV.html.

48. John Taylor, *All the vvorkes of Iohn Taylor the water-poet* (London: John Beale, Elizabeth Allde, Bernard Alsop, and Thomas Fawcet, 1630), STC 23725; ProQuest, *Early English Books Online,* "All the vvorkes of Iohn Taylor the water-poet," https://www.proquest.com/eebo/docview/2264202925/99852944.

49. ProQuest, *Early English Books Online*, "Search syntax and field codes," https://search.proquest.com/help/academic/webframe.html?EEBO_field_codes.html#EEBO_field_codes.html.

50. Text Creation Partnership, "Assigning TYPEs to DIVs," https://textcreationpartnership.org/docs/dox/DIV.html

51. John R. Ladd, "Imaginative Networks: Tracing Connections among Early Modern Book Dedications," *Journal of Cultural Analytics* 3, (2021): 67, 76.

52. Ladd, "Imaginative Networks," 72.

53. The analysis of EEBO-TCP's XML files was performed with the Python programming language, using the ElementTree library. The code used for this study is available at https://github.com/ToTheReader/Computing-Book-Parts.

54. William Patten, *The calender of Scripture* ([London: Richard Jugge, 1575]), STC 19476; A09165.P4.xml.

55. The full distribution data is given alongside the full list of division types at https://zenodo.org/record/4926442#.Yqyn-OzMJyw. (DOI 10.5281/zenodo.4926442).

56. Two sources of genre data were used: the USTC's "subject classifications" which accompany EEBO records as of 2020, and data kindly provided by Alan Farmer and Zachary Lesser, collected for their study "What is Print Popularity? A Map of the Elizabethan Book Trade," in *The Elizabethan Top Ten: Defining Print Popularity in Early Modern England*, ed. Andy Kesson and Emma Smith (Farnham: Ashgate, 2013), 19–54. The language of each text was identified using the fastText library (https://fasttext.cc/).

57. Arthur F. Marotti, *Manuscript, Print, and the English Renaissance Lyric* (Ithaca: Cornell University Press, 1995), 293. On the relationship between dedications and addresses to readers, see Devani Singh, "Dedications, Epistles to the Reader, and Prefatory Custom in Printed English Playbooks, 1559–1642," *The Review of English Studies* 72, (2021): 280–300.

58. Michael Gavin, "How to think about EEBO," *Textual* Cultures 11, (2017): 99; EEBO-TCP, "About the partnership," https://textcreationpartnership.org/about-the-tcp/.

59. Peter C. Herman, "EEBO and Me: An Autobiographical Response to Michael Gavin, 'How to Think About EEBO'," *Textual Cultures* 13, no. 1 (2020): 215.

60. The social and economic contexts and effects of offshore labor are considered, for example, in Jamie Peck, *Offshore: Exploring the Worlds of Global Outsourcing* (Oxford: Oxford University Press, 2017) and B. W. Lambregts, Niels Beerepoot, and Robert Kloosterman, eds., *The Local Impact of Globalization in South and Southeast Asia: Offshore Business Processes in Services Industries* (Abingdon: Routledge, 2016).

61. Text Creation Partnership, "TCP Staff," https://textcreationpartnership.org/about-the-tcp/tcp-staff/.

62. Andie Silva, *The Brand of Print* (Leiden: Brill, 2019), 190.